Enrichment of miRNA targets in REST-regulated genes allows filtering of miRNA target predictions

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

eingereicht an der
Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

von
Dipl.-Biotech. Marie Luise Gebhardt, geb. Sauer

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Lebenswissenschaftlichen Fakultät:
Prof. Dr. Richard Lucius

Gutachter:
1. Prof. Dr. Uwe Ohler
2. Prof. Dr. Miguel Andrade
3. Prof. Dr. Ana Pombo

eingereicht am: 28.09.2015
Tag der mündlichen Prüfung: 15.12.2015

Papi

Abstract

It is very challenging to measure miRNA binding to target 3'UTRs in a genome-wide scale experimentally. Hence, scientists usually make use of computationally generated miRNA-target predictions, which suffer from high false positive rates (24-70%). Since the databases with information on transcription factor binding to DNA from ChIP-sequencing (ChIP-seq) experiments are growing, and since there are sets of genes that are known to be regulated by miRNAs and transcription factors in a coordinated way, we wanted to use knowledge on transcription factor binding to improve miRNA-target predictions.

Due to a huge amount of ChIP-seq data on the transcriptional repressor REST and a proven co-operation with the post-transcriptional level in gene regulation, we chose REST as study object.

First, to be able to make full use of ChIP-seq data, we performed a benchmarking on peak-gene association methods. We found that the choice of a proper peak-gene association method is dependent (i) on the distribution of binding sites of the factor of interest with respect to potential target genes and (ii) on the nature of the follow up experiments to be conducted with the resulting gene lists. Regarding REST, a search for peaks in a window of up to $\pm$10 kb of the transcription start site is appropriate, and the 'ranked' method developed by us can be applied.

To identify miRNAs that co-operate with REST in the regulation of common target genes, we performed a search for over-represented targets of a set of 153 miRNAs, using the predictions of TargetScanHuman 6.2. It was important to develop a random sampling strategy that compensates for biases with impact on the number of miRNA-gene associations found in the analyzed gene set (e.g. 3'UTR length). The algorithm was made publicly available as a web application called 'mBISON'.

Using the developed algorithm we found 20 miRNAs with enriched targets in REST target gene lists from 14 cell types. There is a set of genes, one third of them exhibiting a neural function, that seems to be controlled by REST and the 20 miRNAs by means of varying network motifs to guarantee smooth cellular processes and cell type specificity. During this process, we were able to assign functions to miRNAs and to obtain a global view of the REST-miRNA-target network. We explored the integration of other types of biological data, such as sequence information (motif search), expression and DNase I hypersensitivity data. The procedure was applied to other factors than REST, e.g. to activators, and over-representation of miRNA targets was often found.

We found evidence that our algorithm in conjunction with ChIP-seq data can be used to filter miRNA-target predictions because predicted miRNA-target pairs, that are coordinately regulated by REST and by a miRNA enriched in REST targets, are more likely to be true positives than other pairs.

Zusammenfassung

Eine große Herausforderung der Molekularbiologie besteht darin, miRNA-Bindestellen in 3'-untranslatierten Bereichen von Transkripten auf gesamtgenomischer Basis zu identifizieren. Daher greifen Wissenschaftler für miRNA-Bindestellen gewöhnlich auf Vorhersagen von Computerprogrammen zurück, die aber einen hohen Anteil an falsch-positiven Ergebnissen enthalten (24-70%). Da Datenbanken aus ChIP-Sequenzierung (ChIP-seq) mit Informationen zu Bindestellen von Transkriptionsfaktoren an der DNA stetig wachsen und bekannt ist, dass Transkriptionsfaktoren häufg Gruppen von Genen mit miRNAs gemeinsam regulieren, sollten in der vorliegenden Arbeit Datensätze aus ChIP-seq dazu verwendet werden, Vorhersagen von miRNA-Bindestellen zu filtern.

Es gibt eine große Zahl an ChIP-seq Daten für den Transkriptionsfaktor REST, der nachweislich in der Genregulation mit miRNAs kooperiert und sich folglich für uns als Test-Faktor eignet.

Um die Informationen aus ChIP-seq Daten voll ausschöpfen zu können, wurden zunächst Methoden der Zuordnung von potentiell regulierten Genen zu darin dokumentierten *peaks* getestet. Dabei stellte sich heraus, dass die Wahl der richtigen Methode (i) von der Verteilung der Bindestellen des Transkriptionsfaktors in Bezug auf potentiell regulierte Gene und (ii) von den Folgeexperimenten abhängt, die mit den resultierenden Genlisten durchgeführt werden sollen. Im Falle von REST eignet sich eine Suche von Bindestellen in einem festen Fenster von maximal ± 10 kb Größe *up-* und *down-stream* vom Transcriptionsstart oder die *ranked* Methode, die von uns entwickelt wurde.

Um miRNAs identifizieren zu können, die mit REST bei der Regulation gemeinsamer Zielgene zusammenarbeiten, wurde eine Suche nach überrepräsentierten Zielgenen einer Liste von 153 miRNAs durchgeführt, unter Verwendung der Vorhersagen von TargetScanHuman 6.2. Bei der dafür entwickelten Randomisierungsmethode wurde darauf Wert gelegt Tendenzen zu kompensieren, die unerwünschten Einfluss auf die Anzahl von miRNA-Zielgen-Paaren nehmen (z.B. Unterschiede zwischen Testgenen und Hintergrundgenen in Bezug auf die Länge des 3'-untranslatierten Bereiches). Der Algorithmus wurde der Öffentlichkeit in Form einer Web-Anwendung mit dem Namen ‚mBISON' zugänglich gemacht.

Unter Verwendung des entwickelten Algorithmus wurden 20 miRNAs gefunden, die eine Anreicherung von Zielgenen in den Zielgenen des transkriptionalen Repressors REST (in 14 Zelltypen) aufweisen. Es gibt eine Gruppe von Genen, die zu einem Drittel aus Genen mit neuraler Funktion bestehen, die von REST und den 20 miRNAs mithilfe verschiedener Netzwerk-Motive kontrolliert werden, um reibungslose zelluläre Abläufe und eine Aufrechterhaltung der zellulären Spezifität zu garantieren. Einigen miRNAs konnten bisher unbekannte Funktionen zugeordnet werden, außer-

dem konnte eine globale Sicht auf das REST-miRNA-Netzwerk gewonnen werden. Es wurden andere biologische Daten wie Sequenzinformationen, sowie Expressions- und DNase-I-Hypersensitivitätsdaten integriert. Außerdem wurde das Verfahren auf andere Faktoren als REST, wie z.B. Aktivatoren, angewandt. Überrepräsentation von miRNA-Zielgenen wurde oft gefunden.

Es wurden Hinweise gefunden, dass der Algorithmus in Verbindung mit ChIP-seq Daten zum Filtern von miRNA-Zielgen-Vorhersagen verwendet werden kann, denn miRNA-Zielgenpaare, die gleichzeitig von REST reguliert werden, haben einen größere Wahrscheinlichkeit wirklich zu existieren.

Contents

1 Introduction

1.1 Motivation and Overview

Scientists agonized about the question how microRNA (miRNA) binding to target mRNA can be predicted computationally with high accuracy because for many years a systematic experimental detection was quite challenging (Thomson et al., 2011). Despite the use of sequence features, classification learning, miRNA co-targeting and integration of experimental data in countless approaches, predictions still suffer from false positive rates between 24 and 70% (see Section 1.3.1, Thomson et al., 2011). In contrast, the experimental detection of transcription factor binding has left stages of fledgling due to the development of the Chromatin immuno-precipitation (ChIP) followed by next generation sequencing (ChIP-seq) technique. It facilitates the generation of possible transcription factor target gene lists for the respective experimental condition. Databases that store ChIP-seq data are ever growing and, therefore, provide an unprecedented source of experimental transcription factor binding information.

Demonstrably, a co-operation of transcription factors and miRNAs exists in the regulation of target genes (Shalgi et al., 2007) leading to overlapping target lists. Thus, it is very possible that information on miRNA targets can be derived from the more reliable transcription factor binding information. Accordingly, we hypothesized that information on over-representation of miRNA targets in gene lists obtained from ChIP-seq data could assist in filtering miRNA target predictions. This question will be examined and answered in the present work.

Shalgi et al. (2007) demonstrated that the transcriptional repressor RE1-silencing transcription factor (REST) has targets enriched in certain miRNAs, indicating a strong interaction of both regulatory levels. Together with the fact that a wide base of ChIP-seq data is available for the transcriptional repressor, this knowledge convinced us to choose REST as a study object.

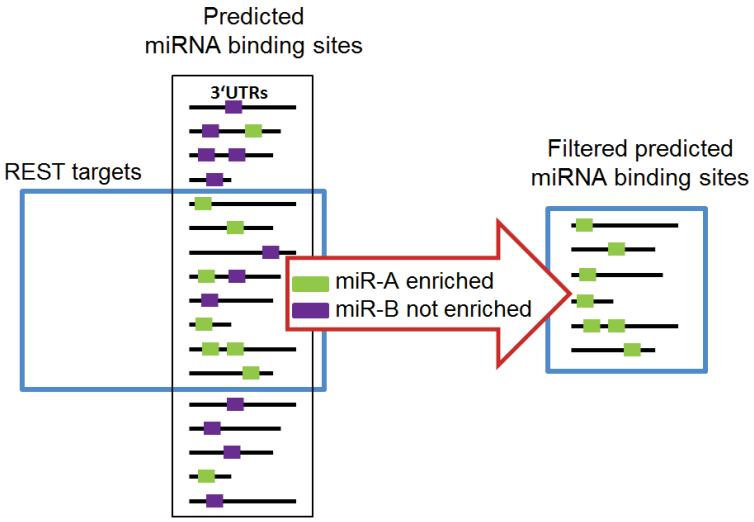

Figure 1.1: Genes co-regulated by REST and by a certain enriched miRNA miR-A comprise a subset with a higher fraction of true positive miRNA target predictions than can be found in the unfiltered predictions. Figure from (Gebhardt et al., 2014).

The general approach is depicted in Figure 1.1. First miRNAs highly integrated in the REST regulatory network will be identified by means of analysis on over-representation of miRNA targets in REST target gene lists. We expect the common targets of REST and the over-represented miRNAs to comprise an elevated amount of true positive predictions for the respective miRNA. If this were true, it would allow us to perform a filtering of miRNA target predictions using ChIP-seq data. Moreover it would provide us insight into the underlying regulatory mechanisms of REST and its co-operating miRNAs.

The thesis at hand consists of three chapters. Chapter 1 comprises an overview on the current knowledge on the transcriptional repressor REST, miRNA binding site prediction tools, the ChIP-seq technique and earlier approaches to enrichment analysis of miRNA targets in gene lists.

In Chapter 2 we searched for a proper method for association of target genes with peaks from ChIP-seq. To achieve this, several methods are compared in a benchmarking procedure.

Chapter 3 comprises a description of how the analysis on over-representation of miRNA targets in REST gene lists was designed. Afterwards, miRNAs with enriched targets are identified from 15 different cell types and it is tested, if the filtering procedure works as

expected. Moreover, the REST-miRNA regulatory network around REST target genes is analyzed and conclusions on miRNA function within this network are drawn. Finally, a web server is presented, in which the search for over-represented miRNA targets is implemented and made available to the scientific community.

1.2 The transcriptional repressor REST and its properties

REST is a transcription factor that is famous for its repressive effect on neural genes in non-neuronal tissue. The factor is well conserved among human, mouse and rat (Palm et al., 1999). In the following section information from the three organisms will be united to give a complete picture of the transcriptional repressor. Information on human will be provided, wherever possible.

1.2.1 Protein structure and expression pattern

According to the UCSC Genome Browser (Kent et al., 2002) the human *REST* gene lies on chromosome 4 and can be transcribed from three different transcription start sites (TSSs). Transcripts of varying length and composition can be generated, which are rearranged to form at least five different splice variants (Palm et al., 1999). The REST protein comes in four isoforms, which are depicted in Figure 1.2, with Isoform 1, generally referred to as 'REST', being the longest and most prevalent. 'REST' is highly expressed in embryonic tissue. The expression level in non-neuronal tissues decreases during differentiation but not as much as in neural tissues. Nevertheless, 'REST' can still be detected in the adult nervous system (Chong et al., 1995; Palm et al., 1998). Isoform 2 has an expression pattern similar to Isoform 1, but it is a very short version of the protein. It retains only zinc finger 1 to 4 and, therefore, exhibits restricted DNA binding capacities. Isoform 3, also known as REST4, is another truncated version of

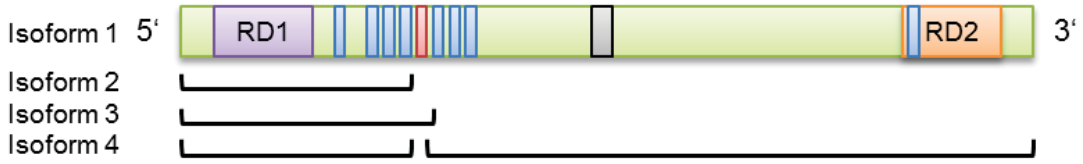

Figure 1.2: Structure of four REST isoforms explained by means of Isoform 1. It has a length of 1,097 amino acids and comprises two repression domains (RD1 and RD2) (Thiel et al., 1998), eight zinc finger motifs in blue and red. Zinc finger 5 (red) exhibits a nuclear localization sequence (NLS) (Palm et al., 1998; Shimojo et al., 2001). A repeat sequence at 512 amino acids is shown in gray (Chong et al., 1995).

Isoform 1 with one zinc finger more than Isoform 2 for translocation to the nucleus (Palm et al., 1998; Shimojo et al., 2001). Isoform 4 (REST5) lacks only the fifth zinc finger of Isoform 1. Isoform 3 and 4 both have a neural-specific expression pattern in embryonic as well as in adult brain (Palm et al., 1998), and regulate target sets different from Isoform 1 (Gillies et al., 2011).

1.2.2 Binding motif and DNA binding profile

REST is a good study object because in contrast to most other transcription factors it binds a comparably well defined, long and non-redundant sequence motif. The 21 bps long motif was found by two scientific groups working independently and almost at the same time. It was designated repressor element 1 (RE1) (Kraner et al., 1992) and neuron-restrictive silencer element (NRSE) (Mori et al., 1992). As a consequence when the transcription factor was characterized in 1995 by two independent groups, the names RE1-silencing transcription factor (Chong et al., 1995) and neuron-restrictive silencer factor (NRSF) (Schoenherr and Anderson, 1995) were chosen.

Over the years a handful of sequence motifs were suggested for the REST binding site based on various combinations of experimental and computational approaches (Bruce et al., 2004, 2009; Johnson et al., 2007, 2006; Otto et al., 2007). A canonical RE1 motif that is bound by REST with high affinity, turned out to control genes that are common to many cell types, while atypical sequence motifs tend to be situated around tissue-specific genes (Bruce et al., 2009). The canonical sequence is separated into two half-sites with a spacer region that can be extended or compressed in respect to the

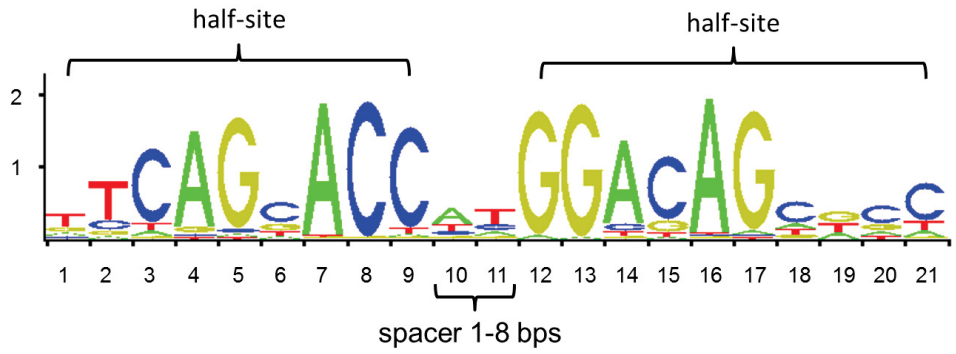

Figure 1.3: Canonical sequence motif of REST binding sites, which can be modified with a spacer region of varying length. Adapted from the JASPAR database (Mathelier et al., 2014).

canonical spacer. Sometimes one half-site is sufficient to enable REST binding (Bruce et al., 2009; Johnson et al., 2007, Figure 1.3).

There is no agreement on the overall distribution of REST binding sites in respect to annotated genes, but all groups found that about 30% of the REST binding sites are located inside the introns of the potential target genes (Johnson et al., 2007, 2006; Otto et al., 2007). Moreover, it is highly likely to find them close to the TSS in range of 2 kb up- or downstream (Arnold et al., 2012).

1.2.3 Function and target genes

Initially REST was found to repress neural targets in non-neuronal tissue (Chen et al., 1998), but its function turned out to be much more versatile. In embryonic stem cells (ESCs) it was suggested to promote self-renewal and pluripotency (Johnson et al., 2008a; Singh et al., 2008). It is a key regulator of neural differentiation (Park et al., 2007), often in close interaction with sets of miRNAs (Conaco et al., 2006). Due to these functions it is consistent to find REST implicated in many different kinds of cancer. It can act as oncogene in tumors from neural cell types because high concentrations of the factor help to assure self-renewal there (Fuller et al., 2005; Kamal et al., 2012; Liang et al., 2014). A low level of expression of a defective variant of REST was found in breast cancer and small cell lung carcinoma (Coulson et al., 2000; Wagoner et al., 2010).

The transcriptional repressor is important in Huntington's disease, epilepsy, Alzheimer's disease and aging (Gillies et al., 2011; Johnson and Buckley, 2009; Lu et al., 2014; Zuccato et al., 2003), and has other functions as regulator of fetal heart development, osteoblast differentiation, splicing events and macroRNAs, to name only some examples (Johnson et al., 2009; Kuwahara, 2013; Liu et al., 2015; Mikulak et al., 2012).

1.2.4 Co-factors and epigenetics

CoREST is a co-factor of REST that is needed for long-term repression of neural-specific target genes. It interacts with the C-terminal zinc finger motif of the repressor (Andres et al., 1999). Another co-repressor is SIN3A, which interacts with the N-terminal repression domain of REST to repress in the promoter region of target genes (Huang et al., 1999). Each of the co-repressors recruits further proteins including histone deacetylases (HDACs), histone H3K4 demethylase LSD1, methyl-CpG-binding protein MeCP2, G9a histone methyltransferase and proteins with chromatin remodeling activity by acetylation as BRG1. By means of the co-factors, in addition to transient repression, REST is also able to perform long-term silencing, where repression remains active even when it

leaves the RE1 site (see Ooi and Wood, 2007, for review). The repression states of the respective target genes are reflected in their histone marks. Genes regulated by REST in many cell types show the lowest expression levels and are dominated by repressive histone marks H3K9me2 and H3K27me3, while genes with more specific repression have patterns of co-existing active and repressive histone marks (Bruce et al., 2009). REST is able to recruit Polycomb complexes to regulate the chromatin state. It was found that this property is independent of the repressive activity of the factor (Arnold et al., 2012; Dietrich et al., 2012).

1.2.5 Regulation of REST expression

REST gene expression is regulated by the pluripotency related transcription factors Nanog, Sox2 and Oct4 (Boyer et al., 2005) and by HIP1 protein interactor, which is involved in Huntington's disease (Datta and Bhattacharyya, 2011). Johnson et al. (2007) suggest a negative auto-regulatory feed-back loop for REST. Some miRNA binding sites (miR-9, miR-29a, miR-153) can be detected in the REST 3' untranslated region (UTR), most of them being REST targets themselves (Wu and Xie, 2006). The neural-specific Ser/Arg repeat-related protein of 100 kDa (nSR100) is a factor involved in splicing that promotes the production of REST4 in neural cells. Since REST4 has a reduced repressive activity, nSR100 indirectly activates the expression of 'REST' target genes. Conversely 'REST' down-regulates nSR100 (Raj et al., 2011).

Collectively, such a number of regulatory interactions and loops suggests that the transcriptional repressor REST is highly integrated into fundamental regulatory networks.

1.2.6 REST-miRNAs

As already indicated in Section 1.2.5, REST regulates several miRNA genes (that we call REST-miRNAs for simplicity hereafter). Quite a few studies have tried to elucidate repressed miRNAs with Johnson and Buckley (2009) probably being the most complete one (see Suppl. Table S2). The authors present about 40 REST-miRNAs, among them famous neural miRNAs such as miR-124 as well as miR-132, which is object of studies related to neural cell death (Hwang et al., 2014; Visvanathan et al., 2007).

1.2.7 Antibodies

When discussing results on Chromatin immuno-precipitation, one has to take into consideration, that different kinds of antibodies for targeting REST exist. Some of them target the N-terminus of the protein (Santa Cruz H290, (Bruce et al., 2009); anti-REST

12C11, (Chen et al., 1998)) and others the C-terminus or internal regions (Upstate 07-579, Santa-Cruz P18, (Bruce et al., 2009)). The list is far from complete. Importantly, antibodies targeting the N-terminus will detect all isoforms of REST.

1.3 miRNAs

miRNAs are the most abundant form of small RNAs with a length of about 22 nucleotides in their mature state. They are transcribed by RNA polymerase II from intronic or exonic sequence of coding or non-coding genes and are often arranged in clusters, where they are co-transcribed from one promoter, gaining pri-miRNAs. Some miRNAs are transcribed in conjunction with host mRNAs of protein coding genes (Bartel, 2004).

miRNA's transcriptional regulation is subject to transcription factors and epigenetic regulators with mechanisms similar to protein coding genes. The processing of pri-miRNAs by the RNase II endonuclease Drosha and DGCR8 in the Microprocessor complex to pre-miRNAs, the subsequent export by Exportin 5 and further cleavage by the RNase II endonuclease Dicer, to release a small RNA duplex with subsequent loading onto an AGO-protein, to build an active RNA-induced silencing complex (RISC), is described in detail in a recent review (Ha and Kim, 2014).

By regulating mRNA decay and translation, animal miRNAs have control over almost all protein-coding genes (Friedman et al., 2009). After incorporation into the RISC complex the passenger strand of the miRNA is removed and the remaining single stranded miRNA is stabilized by the complex (Ha and Kim, 2014). It guides the complex to bind a circular mRNA molecule that is ready for translation, often by imperfect base pairing. The silencing machinery binds the 5'-cap structure and interferes with initiation factors (e.g., eIF4F) for repression of translation or guides its targets to the 5'-to-3'-mRNA decay pathway by promoting first de-adenylation and afterwards removal of the 5'-cap by de-capping enzyme DCP2 and co-factors with subsequent degradation by the 5'-to-3' exonuclease XRN1. Degradation of target mRNAs seems to be the predominant mode of regulation in mammals (see Huntzinger and Izaurralde, 2011, for review).

Although newer findings demonstrate that miRNAs can also perform regulation as activator on post-translational level (Vasudevan, 2012), miRNAs are widely considered to be repressors on post-transcriptional level with involvement in virtually all cellular processes and effects ranging from fine-tuning to significant alterations in expression (Bartel, 2009). In addition, they are often part of feed-forward loops (Tsang et al., 2007)

and perform regulation in a combinatorial and overlapping manner with members of the same miRNA family (Grimson et al., 2007). These two properties are thought to confer robustness to transcriptional programs in case of fluctuations in mRNA expression levels as well as against perturbations from the environment (Ebert and Sharp, 2012; Shalgi et al., 2007; Stark et al., 2005; Tsang et al., 2010). They have, however, profound consequences for scientists, who try to identify specific functions of a miRNA experimentally, e.g. by finding miRNA target mRNAs. Knock-down of a single miRNA rarely leads to detectable phenotypical changes (Miska et al., 2007). Some miRNA families have ten or more members and they have to be deactivated in a concerted way to find out essential functions of the miRNA family (Ventura et al., 2008). Then it is difficult to make statements about the contribution of each member to the phenotype. Ectopic expression experiments are a well established tool that help to determine miRNA function, with the major drawback that they often lead to the identification of false positive targets, for example when the miRNA is over-expressed in a cell type, where it is usually not expressed (Vidigal and Ventura, 2015).

Until recently it has not been possible to detect the direct binding of a certain miRNA to target mRNA in large-scale and systematically (Grosswendt et al., 2014). As a result, scientists often have to rely on computationally found miRNA target predictions. Base pairing to target mRNA is determined by the miRNA seed, 2 to 8 nucleotides from the 5' end of the miRNA, and some downstream nucleotides also contribute to miRNA specificity (Bartel, 2009). Since the seed sequence is very short and imperfect seed-pairing is allowed, prediction of miRNA binding sites is quite error prone (Thomson et al., 2011). The following section will give an overview of attempts that have been made to predict miRNA binding sites in target mRNAs.

1.3.1 Prediction of miRNA targets

The 'simple' approaches to miRNA target identification relied on sequence features of mRNA and miRNA to predict the binding sites of the respective miRNA. Major features that were used, are: seed sequence, mRNA conservation, binding energy between a miRNA and its possible target and co-operative control potential via multiple binding sites. Famous representatives are TargetScan(S), DIANAmicroT, MiRanda, PicTar, PITA and miRWalk using the features in various combinations (Dweep et al., 2011; Enright et al., 2003; Friedman et al., 2009; Kertesz et al., 2007; Kiriakidou et al., 2004; Krek et al., 2005). Most of these methods allow the identification of conserved and non-conserved binding sites, but acceptable precision could only be achieved for conserved binding sites (Selbach et al., 2008). Thus, countless follow-up attempts combined the

already known with new features and with classification techniques such as support vector machines, neural networks or hidden Markov models. The predictions exhibited enhanced precision and specificity primarily for non-conserved target sites (Chandra et al., 2010; Oulas et al., 2012; Sturm et al., 2010).

Beyond the already exploited features, computational miRNA target identification can be improved by integration of experimental data. Very often expression data of mRNAs and/or miRNAs are used (see Naifang et al., 2013, for review), but also KEGG-pathways and Gene Ontology terms (Hsu et al., 2011; Stempor et al., 2012) and data from cross-linking immuno-precipitation (CLIP) can be integrated (Hafner et al., 2010; Liu et al., 2014; Rennie et al., 2014). The improvements can only be done within the frame of the applied experimental conditions; therefore, delivering miRNA target predictions on a genome-wide level is not possible with these methods. The same is true for databases that store experimentally verified miRNA-target interactions (Vergoulis et al., 2012). Thus, if miRNA target predictions on genome scale are needed, a scientist will most likely choose one of the 'simple' approaches (e.g., TargetScan, PicTar or PITA).

According to two studies comparing the performance of the 'simple' methods, TargetScan is a reliable choice. In Figure 1.4 miRNA target predictions are evaluated in two experiments using stable isotope labeling with amino acids in cell culture (SILAC and pSILAC; Beak et al., 2008; Selbach et al., 2008). In both experiments TargetScan was among the best performing miRNA target prediction algorithms.

The TargetScan algorithm

The TargetScan algorithm, as used for predictions from TargetScanHuman 6.2, is based on TargetScanS (Lewis et al., 2005), which considers site conservation and seed matches (see Section 1.5) with either a perfect match of seven nucleotides or with a match of six nucleotides followed by an anchoring Adenin. In addition, the predictions come with a 'context score' calculated from local AU content, co-operativity of sites, distance of residues pairing to the miRNA at positions 13 to 16, and position of the binding site with respect to the stop-codon and the miRNA center. The context score is a measure of efficacy in repression for each conserved or non-conserved miRNA binding site (Grimson et al., 2007).

The manner in which seed sequences are defined leads to overlapping patterns of miRNAs that share six nucleotides in the seed sequences. For broadly conserved miRNA families this leads to six pairs with overlapping target sets (see Suppl. Table S3).

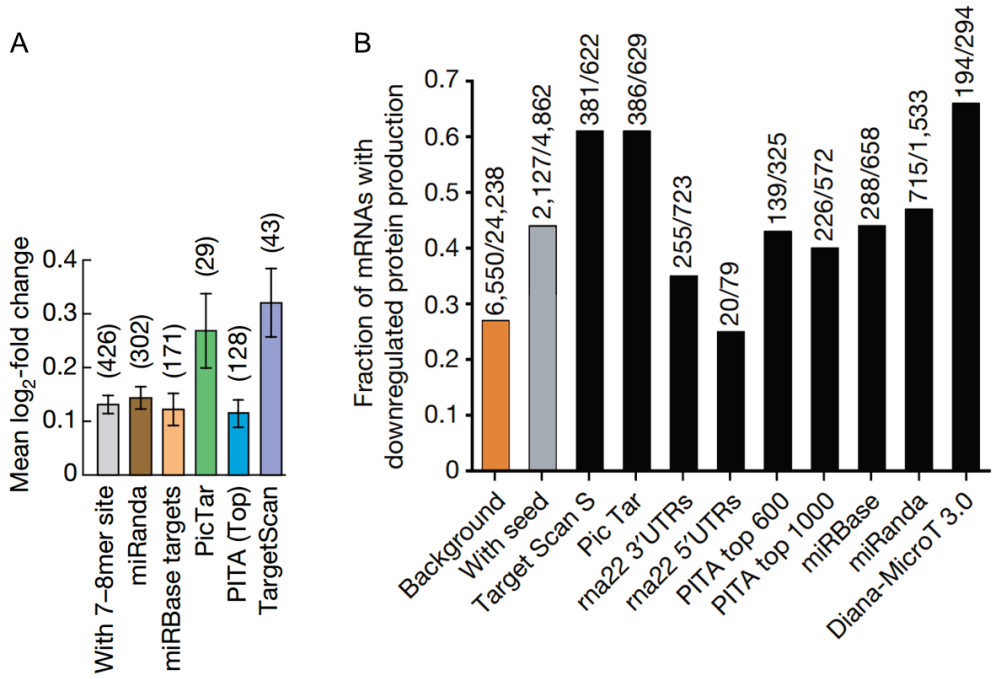

Figure 1.4: miRNA target prediction algorithms are compared using SILAC methods that measure protein levels.
A (Beak et al., 2008): Here programs that consider conservation were compared by average protein de-repression. In parenthesis the number of quantified proteins can be found. Among the compared algorithms TargetScan identified targets with the highest average protein de-repression.
B (Selbach et al., 2008): The study compared miRNA target prediction algorithms by displaying the fraction of mRNAs with log_2-fold change < -0.1, and also showing the numbers of predictions and correctly identified targets. TargetScan, PicTar and DIANAmicroT achieved the highest fractions of correctly predicted proteins. Among these DIANAmicroT had the lowest specificity.

1.3.2 Regulatory modules and loops

If one defines a regulatory module as a set of genes that is co-regulated in a certain condition and executes a common function (Segal et al., 2003), it is easy to imagine the existence of modules co-repressed both by REST and by a set of miRNAs. The functions of regulators within a big regulatory network can be assessed by analysis of network motifs. miRNAs are often part of feed-back and feed-forward loops (Tsang et al., 2007). When REST is now incorporated into these loops, a limited number of circuits with two and three nodes remains possible:

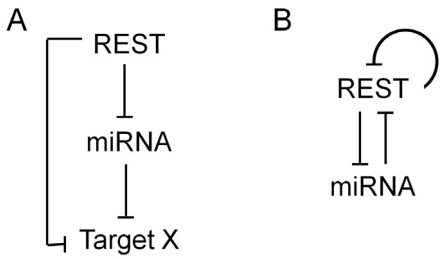

Figure 1.5: Possible network motifs between REST and miRNAs.
A) Incoherent feed-forward loop of type 2 with only repressors as regulators.
B) Negative auto-regulatory feed-back loop of REST and possible double-negative feed-back loop between REST and miRNAs.

Incoherent feed-forward loop of type 2 (I2-FFL): This network motif is very rare in comparison to the incoherent feed-forward loop of type 1 (I1-FFL, containing an activator instead of REST; Alon, 2007), which can be found in many cellular systems (Megraw et al., 2013). The biological function of the I2-FFL is not well understood, but the existence of a I2-FFL for REST and some brain-specific miRNA has been proposed earlier (Tsang et al., 2007). Tsang et al. (2007) hypothesized REST-miRNAs and REST-regulated brain-specific targets are coordinately activated with decreasing REST level during neural development.

Feed-back loops: As mentioned in Section 1.2.5 it is likely that REST is part of an auto-regulatory feed-back loop. According to Alon (2007) this network motif can speed up the response time of gene circuits and it is capable of reducing protein-level variation between cells. Moreover, double-negative feed-back loops are possible between REST and a set of miRNAs (see Section 1.2.5). These would lead to an irreversible mutual exclusion in the expression of one of the loop members (Alon, 2007).

1.3.3 Previous work on over-representation analysis on miRNA targets

Analysis on over-representation of miRNA targets has been done before in at least three different studies.

One was done in the framework of an extensive network analysis of transcription factors and miRNAs. Genes were considered to be targets of a certain transcription factor, when a binding site was found in the gene's promoter region by means of a position-specific scoring matrix and miRNA targets were taken from miRNA binding site predictions from TargetScan and PicTar. In the study, co-occurrence of transcription factors and miRNAs was assessed in two different ways. A cumulative hyper-geometric distribution was used to calculate a p-value for each miRNA-transcription factor pair based on the amount of shared genes c', the occurrence of each regulator alone (m_1, m_2) and the total number of genes in the analysis N (Shalgi et al., 2007). The calculations were made using the following equation (Sudarsanam et al., 2002):

$$P = \sum_{i=c'}^{min(m_1,m_2)} \frac{\binom{m_1}{i}\binom{N-m_1}{m_2-i}}{\binom{N}{m_2}}. \tag{1.1}$$

where i is the summation index.

For the second approach a matrix was generated with the regulated genes as columns and the possible regulators (miRNAs and transcription factors) as rows. For each known regulation there was a '1', otherwise there was a '0'. To generate scores and p-values of co-occurrence of the respective miRNA-transcription factor pairs, the matrix was randomized 1,000 times with a procedure that helped to preserve its degree, meaning the number of targets for each regulator stayed the same as well as the number of times a gene was regulated. The generated matrices were randomized again using edge swapping 100,000 times. For each pair of regulators it was counted in how many of the 1,000 matrices the same or a higher co-occurrence number could be obtained than in the original matrix, with a p-value as result. All analyses were done on human data (Shalgi et al., 2007).

In the second study, over-representation of miRNA target genes was searched for in human annotated gene sets of known function. The method used is called mirBridge. It was implemented in MATLAB® and comprised a sampling strategy, where the authors took biases from 3'UTR-length, conservation and GC-content in consideration. They tried to clear their calculations from the biases by (i) constructing a gene neighborhood

for each gene by means of normalized Euclidean distance between the 3'UTRs and (ii) generating a null distribution for sampling from the respective neighborhood (Tsang et al., 2010, refer to the publication for details). They performed the sampling for each miRNA binding motif from TargetScan separately and calculated p-values for the number of seed matches, the number of genes with the respective seed match, the number of conserved seed matches and the number of seed matches above a certain context score. The p-values were used to compute one q-value for each miRNA binding motif (Tsang et al., 2010).

miTEA is a web application for miRNA target enrichment analysis. The developers made use of a statistical method called minimum-mHG to identify enriched genes in the top of two ranked gene lists. One ranked list is the user's input and comprises a ranked gene list, as generated by differential expression analyses. The second ranked list comes from a miRNA target prediction algorithm of choice. The web application provides p-values and a miRNA network as output with enriched miRNAs as nodes and edges whenever there is an overlap in the targets of two nodes. It supports analysis on human, rat, mouse, drosophila and zebrafish (Steinfeld et al., 2013).

1.4 Experimental detection of transcription factor binding in gene proximity

1.4.1 ChIP-sequencing

In this study data from ChIP-seq experiments are used as source of information on transcription factor binding events, with the goal to identify genes regulated by the respective transcription factor. Figure 1.6 describes the technique and the associated data analysis.

1.4.2 Possible errors, biases and remaining problems of ChIP-seq

ChIP-seq is a mature and widely used technique. Nevertheless, errors and biases can arise from each step in the ChIP-seq pipeline (Figure 1.6).

Dectection of protein binding
The higher the level of enrichment of factor bound DNA-fragments compared to the background, the easier and more reliably true binding events can be identified. Hence,

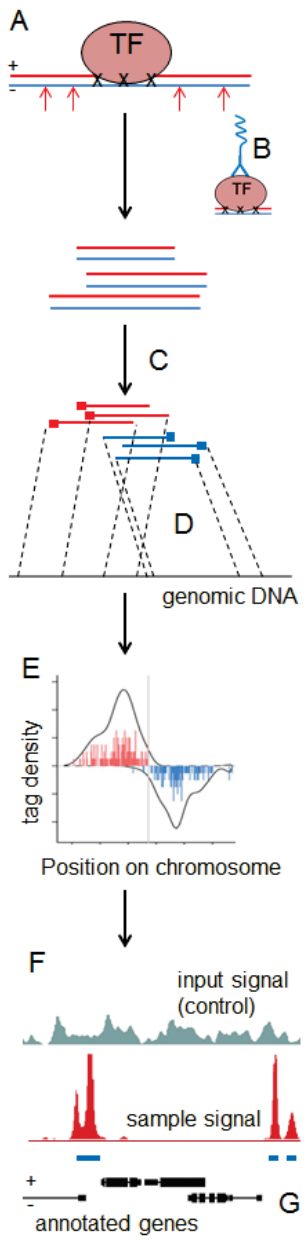

Figure 1.6: ChIP-seq work-flow:
A) Detection of protein binding to DNA *in vivo* by ChIP. Treatment of cellular material under the conditions of interest with formaldehyde leads to cross-linking of proteins to the DNA. Sonication shears the DNA into fragments of 200 to 600 bps.
B) The DNA-fragments that are linked to the protein of interest, are precipitated with a specific antibody and can be reverse cross-linked and purified.
C) Next-generation (or massively parallel) sequencing. The sequencing step results in imaging data that are transformed into sequence-level data by a platform dependent base calling software.
D) Read alignment. Effective alignment of small reads (∼35 bps) to the reference genome in matters of speed, memory, accuracy and flexibility is performed. Short-read mappers allow for mismatches to account for differences to genomic reference sequence originating from sequencing errors, small nucleotide polymorphisms, insertions and deletions, and need strategies for reads from repetitive regions. Three famous examples are ELAND (The Encode Project Consortium I), MAQ (Li et al., 2008) and Bowtie (Langmead et al., 2009), all of them making use of indexing methods.
E) One set of reads for sense and one set for anti-sense are sequenced, leading to a strand-dependent bi-modality in tag density (Wilbanks and Facciotti, 2010).
F) Peak calling. Sequence regions (peaks) with significant enrichment as compared to control or background model are identified. Depending on the peak calling algorithm, regions can be considered as candidate peaks, in which the extended sequence tags a) overlap, b) appear in a fixed clustering distance or c) show up in high counts in fixed width windows across the genome (Wilbanks and Facciotti, 2010, Suppl. Table S4).
G) Peak-gene association. See main text.
The content of this figure description comes from the review (Park, 2009) if no citation is mentioned, Figure adapted from (Kharchenko et al., 2008) and (Park, 2009).

the quality of the applied antibody in matters of specificity and sensitivity is of great importance. Both aspects have to be tested experimentally in advance (Park, 2009).

Even if the experiment is carried out with greatest care, the reads will not be distributed evenly across the genome for multiple reasons. The solubility of DNA is higher in open chromatin regions, therefore the shearing results in a more efficient fragmentation than in the case of heterochromatin (Park, 2009). A selection bias for fragments with high G+C-content originating from fragment size could in part be overcome with improved sequencing library preparation protocols (Quail et al., 2008). However, in general it is recommended to prepare control samples that can be used for normalization during the peak calling step to eliminate biases that stem from the experimental part of the sequence read generation procedure (Park, 2009).

Most often, *input DNA* is used as control sample, which is DNA removed from the sample of interest before the immuno-precipitation step is done. Also *mock IP DNA* from immuno-precipitation without antibody or DNA from immuno-precipitation using an antibody against an unrelated target (e.g. IgG) can be applied. In general, the control samples should be sequenced much deeper to guarantee a representative background distribution of sequence tags across the whole genome (Park, 2009).

Next-generation sequencing

Sequencing errors are to be expected, which accumulate at the end of the reads (Park, 2009). These include simple exchanged bases but also insertions and deletions. The raw error rates of the various sequencing platforms range from below 0.4 to 13 % (Quail et al., 2012). The sequencing errors complicate read alignment to the reference genome.

Read alignment

As described above there are various reasons for discrepancies between the sequenced reads and the reference genome (sequencing errors, SNPs, insertions and deletions, copy number variations). Most tools are able to manage these problems to a certain extent. Nevertheless, there are always reads that cannot be mapped accurately and will be discarded from the analysis. Often only uniquely aligned reads are used for peak calling (Park, 2009).

Peak calling

During the step of peak calling, wherever possible, comparison to a control sample is essential to address problems with repetitive regions and copy number variations (Park, 2009) and for reasons mentioned above (see *Detection of protein binding*).

A peak calling algorithm can only point to the position of the transcription factor binding event with limited accuracy. Algorithms that make use of directionality-scoring methods (Ji et al., 2008; Jothi et al., 2008) report a window containing the TFBS that is narrower than other algorithms. In order to find the exact binding position a subsequent sequence search for TFBSs has to be performed (Wilbanks and Facciotti, 2010).

1.4.3 Peak-gene association

Peaks of ChIP-seq experiments on transcription factor binding can be used to model the binding affinity of the respective factor to each sequence and to find co-operative interactions with other factors (He et al., 2009; Zambelli et al., 2013). Positional weight matrices and consensus binding motifs can be identified (He et al., 2009; Johnson et al., 2007).

Very important and valuable information can be extracted if the peaks can be assigned to genes regulated by the transcription factor. This task is not trivial and can in most cases not be answered accurately without extensive experimental examination. For lack of better approaches, most computational tools look for genes nearest to a peak, taking into account only peaks within a certain range (e.g. 2 kb or 20 kb) off the TSS (Boeva et al., 2012; Heinz et al., 2010; Ji et al., 2008; Shin et al., 2009; Zhu et al., 2010). The methods can be summarized as 'binary' approaches. 'Linear' approaches take into account the distance of a respective ChIP-seq peak to the TSS and give more weight to proximal peaks (Sikora-Wohlfeld et al., 2013).

Ouyang et al. (2009), in addition to the distance feature, make use of peak intensities to assign scores to possible target genes. Intensities g_k of k peaks around a TSS (max 1 Mb) are weighted by an exponential factor, that contains the distance of the peak d_k and a constant d_0, and are summed up to a value of association strength a_{ij} of a transcription factor j and a gene i.

$$a_{ij} = \sum_k g_k e^{\frac{-d_k}{d_0}} \tag{1.2}$$

It is up to the user to choose the constant d_0. It should be smaller for factors that tend to bind close to the TSS and bigger for factors that bind further away. Default is 5 kb.

Cheng et al. (2011) concentrate on the area around the TSS as well (default width is 10 kb), but they try to take the specific binding behavior of each transcription factor into

account. They call their approach "target identification from profiles" (TIP). Depending on properties such as the number, distance and distribution of binding sites (extracted from ChIP-seq data) a characteristic binding profile is calculated. It is used to rank the potential target genes.

Sikora-Wohlfeld et al. (2013) presented a comparison of the above mentioned methods to a new approach, 'ClosestGene'. In this approach, the distance of a gene to all peaks around the TSS [+/-1 Mb] forms a peak-to-gene distance distribution that is used to score the peaks according to their likelihood of targeting its nearest gene. In its publication, 'ClosestGene' outperforms the above mentioned methods.

1.4.4 Possible errors, biases and remaining problems of peak-gene association

The 'nearest gene' strategy works well for factors that tend to bind in the promoter region of target genes. Nevertheless, to choose always the nearest gene as target can cause a selection bias. If transcription factor binding sites can be found in large intergenic regions more often by chance than in short regions, then genes situated in the neighborhood of large intergenic regions will more often be detected as nearest gene to transcription factor binding sites than genes with flanking coding regions (Taher and Ovcharenko, 2009).

Sandmann et al. (2006) integrated the distances of Mef2 binding sites to a gene with differential gene expression data from a wild type fly and a *Mef2*-loss-of-function mutant[1]. Following a similar principle, scientists often make use of gene expression data from transcription factor perturbation experiments, or even of publicly available expression data, if no accompanying gene expression dataset exists (Wu and Ji, 2013). This procedure identifies direct transcription factor targets with a high likelihood and, therefore, can improve peak-gene association accuracy. It misses TFBSs that are bound by the analyzed transcription factor without having impact on gene expression under the given conditions. Using only binding sites containing the desired transcription factor binding motif can filter false positive peaks. However, one might miss sites with *de novo* binding motifs e.g. half motifs, which occur when the transcription factor dimerizes with another DNA-binding protein.

If a peak lies in proximity to more than one gene, without experimental examination the researcher cannot know if the TFBS regulates one, two, all or none of the respective

[1] In this experiment the information on transcription factor binding stems from the ChIP-chip technique. After amplification and denaturation of the Chromatin immuno-precipitated DNA fragments there is no sequencing step. Instead the fragments are labeled with fluorescent tags and ligated to a microarray of single-stranded DNA probes covering selected genomic positions (Ren et al., 2000).

genes. A decision on the most likely scenario can be facilitated by annotating the peak with respect to genomic features such as promoter, intron, exon, 3'ends and so on.

There are transcription factors that bind in large intergenic regions (silencers, enhancers, insulators) and have impact on gene expression by long-range interactions to promoters (Soler et al., 2010). One example is C/EBPα, which binds a distal enhancer of the human *PU.1* gene, situated 14 kb upstream of the TSS. PU.1 is a transcription factor crucial for myeloid and early B-cell development (Yeamans et al., 2007). The enhancer regulated gene is not in every case the closest located one (Lettice et al., 2003). It is almost impossible to assign peaks from very distant regions to the corresponding target genes correctly without performing further experiments such as chromosome conformation capture assays (3C, 4C, 5C and Hi-C, see Belmont, 2014, for review).

This problem has not been solved up to now. One attempt to address the association of TFBSs in enhancer regions with genes was undertaken by Rodelsperger et al. (2011), who classified nearest genes (up to 2 Mb) by means of conserved synteny, functional similarity to the transcription factor and proximity to the transcription factor in the protein-protein-interaction network. A recall of 58% was achieved (Rodelsperger et al., 2011), but the method has to be adapted to every new dataset. There is no tool to assist in this procedure and thus, it turns out to be not suitable for a web-application or any analysis involving many different datasets.

In summary, one can say that there is no gold-standard way for peak-gene association with ChIP-seq peaks. The method of choice depends on the binding behavior of the analyzed transcription factor and should be evaluated in the given context.

1.5 Definitions and Abbreviations

seed match
The terms 'seed sequence' or 'seed match' designate the up to 8 bps long part of the target mRNA that are recognized by the RISC-complex containing the regulating miRNA (Lewis et al., 2005).

miRNA seed
The miRNA seed are the nucleotides 2-8 of the mature miRNA that are used to guide the RISC-complex (Lewis et al., 2005).

miRNA family
miRNAs with identical miRNA seeds at nucleotides 2-8 of the mature miRNA are called 'miRNA family' (Bartel, 2009).

miRNA family names
For sake of simplicity miRNA family names are abbreviated to 'miR' and the first number mentioned in the family name. For example:
let-7/98/4458/4500 $\Rightarrow$ let-7.
miRNA families will be called miRNAs for sake of simplicity.

REST-miRNAs
miRNAs regulated by REST will be called 'REST-miRNAs'.

Enrichment miRNAs
miRNA families with over-represented targets in REST gene lists will often be referred to as 'enrichment miRNAs'.

Conserved and broadly conserved miRNA families
According to Friedman et al. (2009) *conserved* miRNA families are conserved across most placental mammals, while *broadly conserved* miRNA families are conserved across most vertebrates to zebrafish.

miRNA binding site conservation
The researchers who developed the TargetScan algorithm differentiate miRNA binding sites into *conserved* and *non-conserved*. Site conservation is defined by comparison to 28 vertebrate genomes by conserved branch length as applied for the UCSC Genome Browser (Karolchik et al., 2008). Depending on the site type they define a different threshold t for conservation: 8mer $t \geq 0.8$, 7mer-m8 $t \geq 1.3$, 7mer-1A $t \geq 1.6$.

Table 1.1: Abbreviations

Abbreviation	Term
bp/bps	base pair/base pairs
CDS	coding start
CDE	coding end
ChIP	chromatin immuno-precipitation
ChIP-seq	ChIP-sequencing = chromatin immuno-precipitation followed by next-generation sequencing
CLIP	cross-linking immuno-precipitation
Cl.	cluster
DHS	DNase I hypersensitive
DMEM	Dulbecco's modified Eagle's medium
ESC	embryonic stem cell
FC	fold change
FDR	false discovery rate
GEO	Gene Expression Omnibus
HAIB	HudsonAlpha Institute for Biotechnology
IP	immuno-precipitation
I1-FFL	incoherent feed-forward loop of type 1
I2-FFL	incoherent feed-forward loop of type 2
miRNA	microRNA
NCBI	National Center for Biotechnology Information
NLS	nuclear localization sequence
NP	neural progenitor
NRSE	neuron-restrictive silencer element
NRSF	neuron-restrictive silencer factor
RD	repression domain
REST	RE1-silencing transcription factor
RE1	repressor element 1
RISC	RNA-induced silencing complex
RNA-seq	RNA-sequencing
SILAC	stable isotope labeling with amino acids
SNP	single nucleotide polymorphism
TSS	transcription start site
TES	transcription end site
TFBS	transcription factor binding site
TIP	target gene identification from profiles
UTR	untranslated region

2 From ChIP-seq to gene lists

2.1 Motivation - Choosing the appropriate peak-gene association method

One of the goals of our study is to make use of data from ChIP-seq to gain knowledge about miRNA targets and functionality. A crucial step is to assign potentially regulated target genes to ChIP-seq peaks. In 2010, when this project was initiated, many of the methods mentioned in Section 1.4.4 were not available. Having basically the option to use a 'binary' approach, we wondered whether peak-gene association could be improved by allocating a rank to the genomic feature, in which a ChIP-seq peak is situated. For instance, given a peak placed in the neighbourhood of two genes A and B, if the peak overlaps the promoter of gene A and is just down-stream gene B, one could assume that the binding detected will be having an effect on gene A rather than on gene B. A peak-gene association method could make use of this knowledge to enhance its performance.

The next chapters comprise a comparison of this 'ranked' approach to the simple 'binary' method. The comparison is presented together with the results of newer approaches (see Section 1.4.4) to assess if they can provide significant improvements.

2.2 Methods

2.2.1 Data sources

The transcriptional repressor REST

Arnold et al. (2012) made possible binding positions of REST from ChIP-seq available in conjunction with data on differential mRNA expression before and after knock-out of the transcriptional repressor. The data were collected in mouse ESCs and neural progenitors (NPs) and can be downloaded from Gene Expression Omnibus (GEO, Suppl. Table S1).

Transcripts with a fold change (FC) of at least 1.5 and a false discovery rate (FDR) smaller than 0.1 were considered to be significantly up-regulated in knock-out cells in comparison to the wild type providing a list of genes that were regarded as true positives.

ChIP-seq data were unaligned and in SRA-format. BED-formated files were generated as described in Suppl. Methods *General Methods*.

All methods for peak-gene association described below were tested, and the number of true positives found by each approach was plotted as fraction of all true positives (sensitivity) and of all genes assigned to the REST binding positions (precision).

The transcription factor Androgen Receptor

In the study of Zhu et al. (2012) the authors published ChIP-seq data on Androgen Receptor binding. They specified directly activated and repressed target genes of the transcription factor, which were joined to generate a true positive set (195 activated, 306 repressed, 501 in total). In their experiment, the Androgen Receptor was stimulated by the Androgen Receptor agonist metribolone (R1881) in a cellular model for prostate cancer. Zhu et al. (2012) did a ChIP-seq analysis available as SRA raw data on GEO (Suppl. Table S1). BED-formated files were generated as described in Suppl. Methods *General Methods*. We calculated sensitivity and precision for all peak-gene association methods.

2.2.2 The ranked method

Based on the classic model of a gene, we made the assumption that a transcription factor will most likely perform direct gene regulation if it binds to the promoter region. We consulted the MPromDb database of computationally predicted and known active RNA-Polymerase II promoters for human and mouse (Gupta et al., 2011). For the peak-gene association, all genes with ChIP-seq peaks in these genomic locations were listed (rank 1, Figure 2.1). After removing the respective peaks, the subsequent query was performed

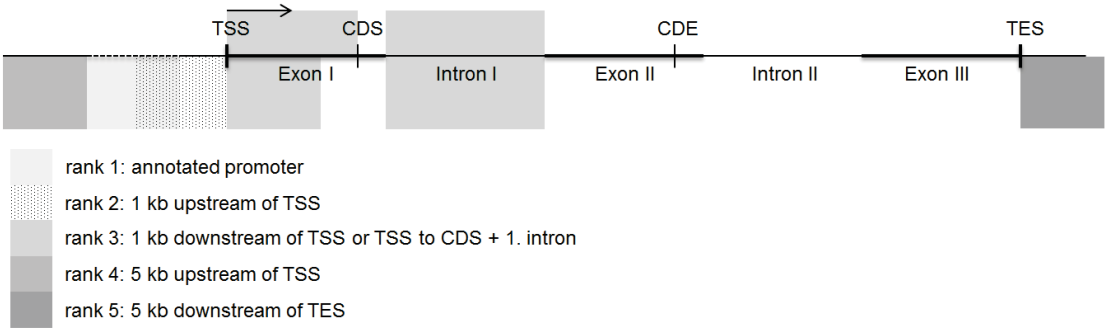

Figure 2.1: Stylized depiction of the five regions used for peak-gene association via ranked method and their priority.

on an area of 1 kb up-stream of the TSS, which was very likely to contain a promoter (rank 2). Genes assigned to one of the peaks were recorded and the peaks were removed as done before.

The procedure was continued with three further regions: 1 kb down-stream of the TSS or (if more than one exon existed) TSS to CDS and the first intron (rank 3), 5 kb up-stream of the TSS (rank 4) and 5 kb down-stream of the transcription end site (TES, rank 5), resulting in a final gene set from the five combined lists.

2.2.3 Benchmarking peak-gene association methods

Several peak-gene association methods were tested for assignment of genes to the transcription factor binding positions.

The scientists who designed the 'ClosestGene' method, made an R-package available that allows the user to easily run multiple peak-gene association analyses with various parameters on one dataset (Sikora-Wohlfeld et al., 2013). It provides scores for every method (1.-4.).

1. Binary method: For this method the scores are just 0 for no peak and 1 if a gene has a peak in a fixed window size. We chose the window sizes to be 1, 2, 5, 10, 20 or 50 kb.

2. Linear method: The 'linear' method was run with default parameters. The method has positive scores as output, with higher values of the score pointing to higher likelihood of obtaining true targets.

3. Ouyang method: For this method the parameter d_0 can be chosen. It is a constant that helps to weight the distance between the peak and the TSS. Due to the fact that REST often binds within 2 kb from the TSS, we chose $d_0 = 2$ kb. For the analysis of binding peaks from Androgen Receptor we chose the default of $d_0 = 5$ kb. The method has positive scores as output, with higher values of the score pointing to higher likelihood of obtaining true targets.

4. ClosestGene method: The 'ClosestGene' method was run with default parameters. We chose z-score as output. It yields positive scores, with higher values of the score pointing to a higher likelihood of obtaining true targets.

5. Ranked method (our method): Experimental Promoter » Promoter 1 kb » first intron » 5 kb up-stream of TSS » 5 kb down-stream of TES.

For comparison we tried two modifications of the ranked method:

'strict': The method was performed as the 'ranked' method without using the fifth rank (5 kb down-stream of TES)

'non-ranked': This method serves as reference for the 'ranked' method. Already assigned

peaks are not removed from the list of ChIP-seq peaks. Consequently, the peaks can be assigned to more than one gene. Using the example from the introduction, if a peak is found in the promoter region of gene A and down-stream of gene B, it is assigned to both genes.

The overlap of genomic regions with RefSeq genes was detected as described in Suppl. Methods *General Methods*.

6. TIP method: The authors of this work provided an R-function and example files that made it possible to convert all data into the required formats. The method is described in depth in the corresponding publication (Cheng et al., 2011).

2.3 Results - Comparison of peak-gene association methods

2.3.1 The transcriptional repressor REST

We compared a list of (likely) true REST target genes from NPs to gene lists generated by several peak-gene association methods, to find out which approach allows to identify high numbers of true targets at low costs of false positives.

The 'binary' approach precision correlated negatively with sensitivity in this experiment (Figure 2.2). It achieved high results in precision from 1 to 5 kb with 50.9 to 35.7% but with sensitivities not higher than 25%. Precision and sensitivity were almost even in a range of 22 to 25% when a window size of 10 kb is used and for all 'ranked' methods. The 'strict' method had the highest precision, 'non-ranked' the highest sensitivity.

None of the newer approaches clearly outperformed the simple 'binary' method (Figure 2.2 and Suppl. Table S5). The results of the 'linear' approach lay roughly on a par with those of 'binary' of 10 and 20 kb window size. Considering targets with scores higher than 0.8 and 0.9 the method achieved a good compromise between sensitivity and precision. The 'Ouyang' method had a high precision. When the cutoff was not chosen too strictly (in range of 0.1 to 1), results became comparable to those of the 'binary' and 'ranked' methods. The performance of the newest method 'ClosestGene' was disappointing. Precisions higher than 35% could be obtained with cutoff 4 but on high costs of sensitivity. The TIP method turns out to be not suitable for peak-gene association of ChIP-seq peaks from experiments on REST binding due to a very low sensitivity.

A very similar result was generated with a list of REST target genes in ESCs (Suppl. Figure S1), with the difference that precision and sensitivity values were generally lower.

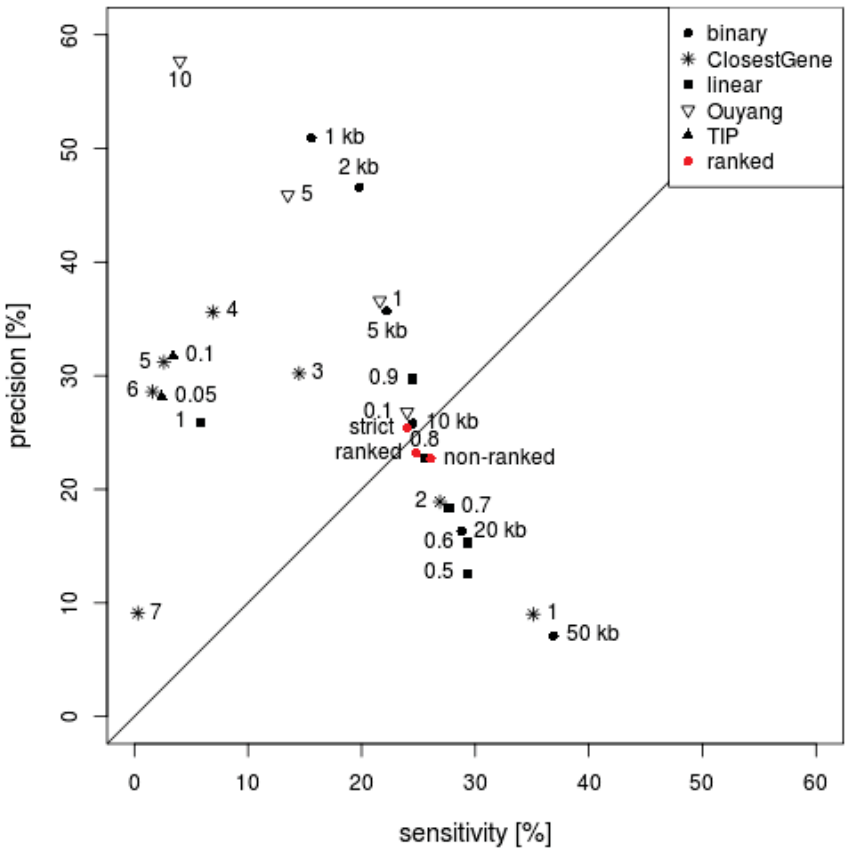

Figure 2.2: Comparison of peak-gene association methods. Precision is plotted against sensitivity. Genes were assigned to ChIP-seq peaks using the example of REST NPs. 379 genes were up-regulated after knock-out of REST in respect to the wild type (point of reference for sensitivity).
Data point labels: Binary - targets in range of 1 to 50 kb window size. ClosestGene - targets with score higher than 1, 2, 3, 4, 5, 6 and 7. Linear - targets with score higher than 0.5, 0.6, 0.7, 0.8, 0.9 and 1. Ouyang - targets with score higher than 0.1, 1, 5 and 10. TIP - targets with p-value smaller than 0.05 and 0.1.

REST binding profile

The 'TIP' method generates a weighted binding profile of the transcription factor before it uses the weights to score target genes. By plotting the weight over the region around the TSS it is possible to have a look at the REST binding profile for the ESC and NP datasets. Figure 2.3 displays two profiles from two replicates each. They differ a lot. For two of the profiles, which do not originate from the same cell type, there is a high peak between 0 and 2 kb down-stream of the TSS. In addition there is a very high peak

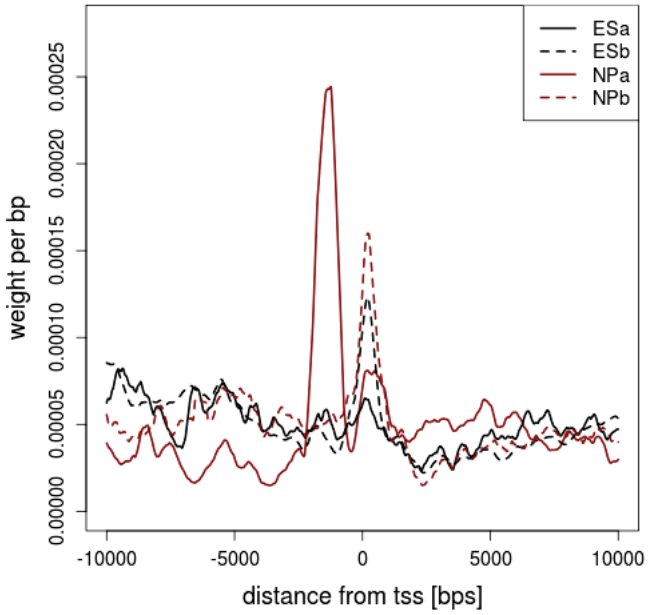

Figure 2.3: REST binding profiles of two replicates for ESCs and NPs. Weights were generated by the TIP algorithm with default parameters from ChIP-seq data.

for one of the NP replicates between 2 kb up-stream of the TSS and 0.

True positive lists: The 457 de-repressed genes in ESCs have 91 genes in common with the 379 genes up after knock-out in NPs. *ChIP-seq gene lists*: The ChIP-seq gene lists of NPs are almost a subset of the ESC gene lists (always $>$ 98 % of NP genes are contained in ESC genes). In general, the genes found by peak-gene association always had a higher intersection with the true positive list of the NPs than with that of the ESCs.

2.3.2 The transcription factor Androgen Receptor

Prostate cancer cells were stimulated with an Androgen Receptor agonist R1881. Afterwards binding of Androgen Receptor was detected by a ChIP-seq experiment. From the resulting list of Androgen Receptor peaks targets were called using six different peak-gene association methods. Figure 2.4 shows the precision and sensitivity that was found for the approaches at various cutoffs.

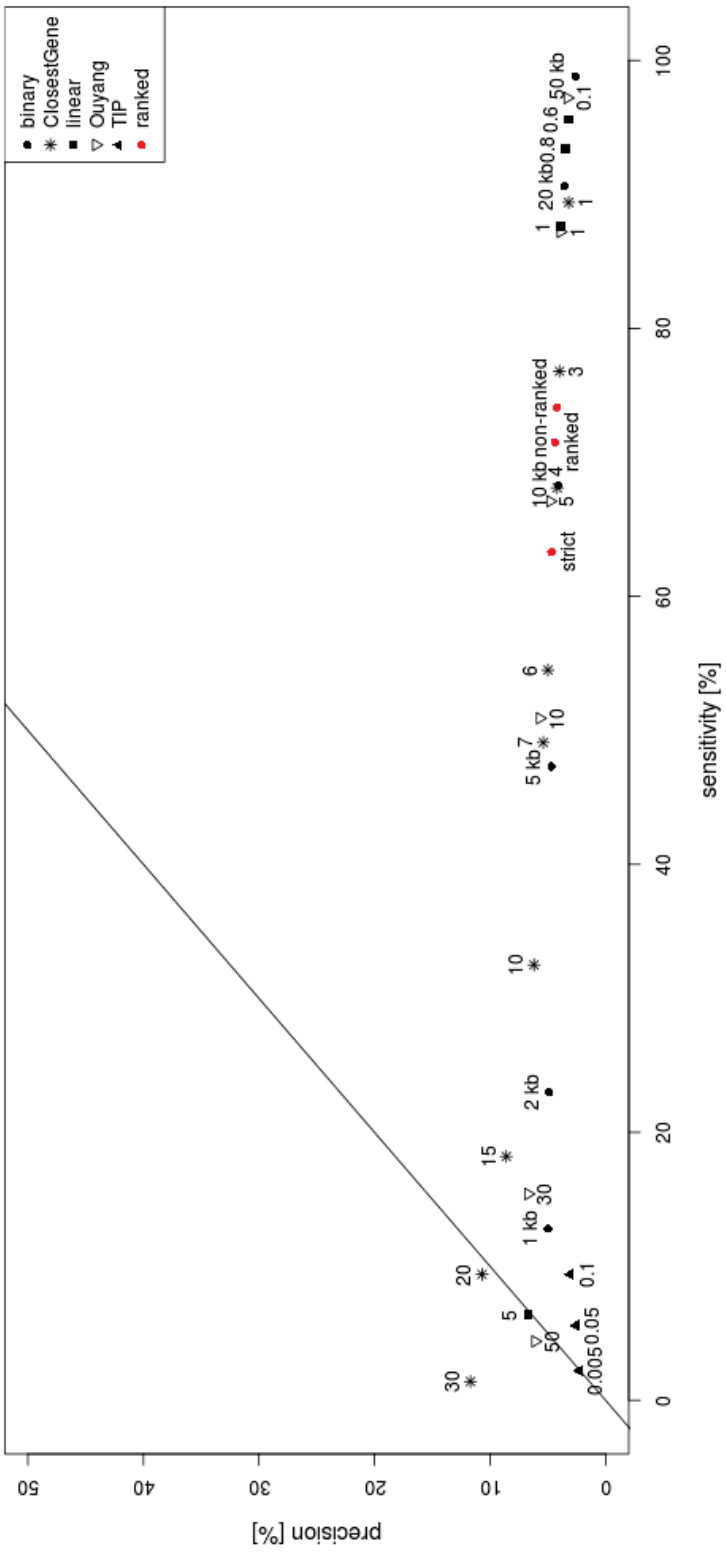

Figure 2.4: Comparison of peak-gene association methods in terms of precision and sensitivity using the example of Androgen Receptor in a prostate cancer model. The transcription factor was immuno-precipitated with the chromatin under stimulated (R1881+) conditions. 501 genes were up- or down-regulated after inhibition by the Androgen Receptor antagonist compound 30 in respect to the wild type (point of reference for sensitivity).
Data point labels: Binary - targets in range of 1 to 50 kb window size. ClosestGene - targets with score higher than 1, 2, 3, 4, 5, 6 and 7. Linear - targets with score higher than 0.6, 0.8, 1 and 5. Ouyang - targets with score higher than 0.1, 1, 5, 10, 30 and 50. TIP - targets with p-value smaller than 0.005, 0.05 and 0.1.

The precision of all peak-gene association methods was low in comparison to the findings in the case of the transcriptional repressor REST. Precision values of more than 10% were obtained only by the 'ClosestGene' method. The performance, however, is very dependent on the cutoff choice. All methods except for 'TIP' allow the identification of Androgen Receptor targets with a high sensitivity, when proper cutoffs and parameters are chosen, but the precision is far below 10% for all of them. The binding profile of Androgen Receptor within 50 kb off the TSS, generated by 'TIP', reveals no obvious binding preferences (see Suppl. Figure S2).

2.4 Discussion

2.4.1 Transcriptional repressor REST

To address the question of which peak-gene association method is the most suitable for our transcription factor REST, we performed the benchmark analysis on varying peak-gene association methods presented in the previous section.

In accordance with former studies and with the weighted REST binding profile from the 'TIP' analysis very high precisions could be achieved when an area of 1 to 2 kb around the TSS was searched by using the 'binary' approach. REST often binds the promoter region of genes. The 'Ouyang' and the 'linear' method score genes higher, when they have a shorter distance to the respective peak. This procedure is favorable in the case of a factor tending to bind close to the TSS. As a result, both methods show a good performance.

The 'binary' method with 10 kb window size behaves similarly to the 'ranked' approaches. This is consistent because in the 'ranked' methods at least 1 kb down-stream of the TSS and the first intron are searched, in addition to 5 kb up-stream of the TSS, summing up to an area of comparable size to 10 kb. There is no striking difference between the three 'ranked' approaches. The ranking of target genes slightly improves the precision (compare 'ranked' and 'non-ranked') at the cost of sensitivity. It more often happens that one binding site has impact on more than one gene than we had expected. In terms of sensitivity, the 'ranked' method does not perform much better than 'strict'. Hence, in the case of REST it will not be wise to include 5 kb down-stream of the TES into the query for most applications. Moreover, the experiments show, that while target calling can still bring satisfying results for approaches that search within 10 kb around the TSS, 20 kb or more are not suitable for REST. This has often been done in the past (Johnson et al., 2007, 2006; Otto et al., 2007).

There is no outstanding method that brings highest precision and sensitivity. However,

we show that the method of choice does not only depend on the binding behavior of the transcription factor but also on the nature of experiments that are to be performed with the resulting gene list.

For an analysis on over-representation of miRNA targets in gene lists it is desirable to have a gene list with only REST target genes, but the list should not be too short. False positive genes can be buffered by the comparison to a random background; therefore, it is possible to choose a method close to the diagonal line (precision = sensitivity) to achieve a compromise in precision and sensitivity. This would mean choosing the 'binary' method with 5 or 10 kb window size, the 'ranked' method and their 'strict' variant, the 'Ouyang' method with $d_0 = 2$ kb and cutoff 0.1 or 1, or the 'linear' method with a cutoff of 0.9 or 0.8. Making the decision for the correct cutoffs is not easy without extensive benchmarking. Regarding the 'Ouyang' and 'binary' method a window size of 2 kb can be guessed correctly due to current knowledge about the transcriptional repressor. However, if simplicity is preferred, the 'binary' method is probably the best choice.

One has to keep in mind that the analysis was performed on ChIP-seq data from experiments in mouse. Results might be different for REST in human. Nevertheless, since the transcription factor is well conserved, we assume that these results can be transfered to human ChIP-seq data.

The experiments above have some major limitations, which have to be mentioned. They are only a statement on active REST. A knock-out experiment can only show transcripts de-repressed that had been actively repressed before. It is known that REST is not active on all binding sites and that it needs co-factors for repression (see Sections 1.2.2 and 1.2.4). As a result, a ChIP-seq analysis on REST will yield a number of peaks close to genes that are true targets in another condition, but which cannot be monitored by the knock-out experiment. This explains why precision values cannot be 100% in this experiment.

In addition, differential expression analysis is not able to distinguish between direct and non-direct targets. Genes in the true positive list with their expression changing upon knock-out of *REST* that are not direct REST targets, have negative impact on the measurements of performance. This might explain why in ESCs both precision and sensitivity were much lower than in NPs. Apparently the NP true positive list contains a much higher amount of directly regulated REST targets than the ESC true positive list, which is shown by the high intersection of the NP true positives with the ChIP-seq gene lists from both ESCs and NPs. It is obvious that in ESCs other or additional pathways are active.

The disagreement of the weighted profiles of the ESC and NP replicates shows that

either the biological conditions were difficult to rerun, or that the ChIP-seq technique was not able to track REST binding correctly. In any case, the benchmarking was set up in a way that only targets were listed, which were detected in both replicates. This negatively affects methods such as 'TIP' and 'ClosestGene' that rely on the binding profile, detect different targets for the replicates and find only a small intersection in the end.

Moreover, one has to keep major limitations in mind that originate from the ChIP-seq data production and can be found in Section 1.4.1. For example, the specificity of the antibody has to be taken into consideration. The applied antibody (Santa Cruz, H-290) is directed against the N-terminus of the REST protein; it captures all its isoforms.

2.4.2 The transcription factor Androgen Receptor

Looking at a second transcription factor, we monitored the performance of various peak-gene association methods on ChIP-seq Androgen Receptor peaks from an experiment with stimulated activity of Androgen Receptor in a prostate cancer model.

The precision level is rather low in general. It can be expected that the list of directly regulated target genes is far from complete. Nevertheless, a comparison of the approaches is possible.

Again the 'TIP' approach fails to achieve high precision and sensitivity values. This is probably due to the fact that the Androgen Receptor binding profile does not deliver clues about a systematic binding behavior.

Zhu et al. (2012) identified 25 kb around Androgen Receptor regulated genes as a range where Androgen Receptor binding sites appear more often than by random expectation. When genes are assigned e.g. by the 'binary' method within a range of 20 or 50 kb almost all genes can be called, but too many false positive genes will be among the targets in the end. When we define the goal to achieve a compromise between precision and sensitivity again and use the diagonal line as reference, the 'ClosestGene' methods with cutoff 15 or 20 would be the method of choice for target calling. It compiles a distribution of peak-to-gene distances for all peaks within 1 Mb of a TSS, pooled for all considered TSS, to be made use of for gene calling. The procedure seems to be favorable in the case of the Androgen Receptor, but the decision on the proper peak-gene association method again depends on the nature of the follow-up experiments.

2.5 Conclusion

In summary the 'ranked' method and the 'strict' variant do not perform worse or better than the comparable 'binary' approaches and it is possible to use them for peak-gene association in the case of REST and factors with comparable binding behavior. There are factors, however, with complicated binding profiles, where it is helpful to make use of newer approaches, e.g. the 'ClosestGene' method. The two very different examples show how important the application of a proper peak-gene association method can be. Any online tool that makes use of peak-gene association should offer a variety of methods to choose for the user.

For most scientists the choice of a proper method for peak-gene association is not part of their basic interests in research. They want to have methods that call target genes correctly without the need of complicated and time consuming benchmarking procedures. Apart from the difficulty to make the right choice on the peak-gene association method, it is always challenging to choose parameters correctly. This is why the developers of the 'ClosestGene' method tried to invent a parameter-free approach. However, we showed that the choice of the score cutoff is of fundamental importance for the outcome of the target calling procedure, again leaving the decision up to the scientist and requiring a benchmark.

A tool for simple benchmarking could be offered with ChIP-seq and differential expression data or a list of true positives as input. It is conceivable to assist in the choice of a proper peak-gene association method using a support vector machine or a neuronal network on basis of a transcription factor binding profile. Such a profile could be generated from a ChIP-seq experiment. Transcription factors or complexes with impact on transcription could be classified according to their binding behavior. Since nowadays the production of ChIP-seq data often runs hand in hand with RNA-sequencing, expression data could be integrated as well, with the final goal of not only recommending a peak-gene association approach, but also of predicting the most suitable parameters and cutoffs.

2.6 Contributions

I did all the computational analyses and interpretation of the results in this chapter under the supervision and with the support of Prof. Miguel Andrade.

3 Analysis on over-representation of miRNA targets in gene lists

3.1 Motivation

From Chapter 2 we know that we can use our 'ranked' method to associate target genes to REST peaks from ChIP-seq data. We did this for 15 human cell types and used the resulting gene lists to answer the question whether ChIP-seq data can be used to filter miRNA target predictions by first identifying miRNAs with over-represented targets in the REST target gene lists, and then determining if the common targets of REST and the respective miRNAs are enriched in true miRNA target predictions.

Knowing that REST-binding close to a gene, even if it has been detected correctly, does not necessarily mean that the gene is a target of the transcription factor, we will refer to these 'potential targets' as 'targets', 'target genes' or 'REST-bound genes' to simplify the following statements and assumptions.

3.2 Methods

3.2.1 Main data sources

ChIP-seq data of REST binding in 15 human cell types

In 2012 the ENCODE project released DNA binding data of the transcriptional repressor REST in 15 different cell types from ChIP-seq experiments with two replicates each in broadPeak-format (Suppl. Table S1). The ENCODE broadPeak-format is an extended BED-format, which contains information about the location of peaks identified during the experiment (Kent et al., 2002). To demonstrate the application of the ranked method (see Section 2.2.2), it was applied to all detected peaks to generate lists of target genes of REST. Only genes identified from both replicates were listed and one set of genes was collected for each cell type. For the data evaluation one has to keep in mind that

the antibody used (anti-REST 12C11) is not able to distinguish the isoforms from the protein.

For some tests an additional dataset was used, which was produced from ChIP-seq on REST in Jurkat T cells by Johnson et al. (2007). The scientists provided their processed data in a supplementary file of their publication.

TargetScanHuman 6.2

miRNA binding site predictions were taken from the TargetScanHuman 6.2 resource (see Suppl. Table S1) for reasons explained in Section 1.3.1. Conserved miRNA binding site predictions for broadly conserved miRNA families (see Section 1.5 *Definitions*) ensured a high accuracy and a focus on few miRNA-gene pairs, which is favorable for the subsequent simulation in terms of computational costs. The downloaded dataset comprised predictions for the longest 3'UTRs of 22,018 annotated human RefSeq genes from UCSC whole-genome alignments. The miRNA binding sites were pooled for all variants associated with each gene. After pooling, 72,770 unique miRNA-gene pairs from 11,161 genes, according to NCBI Entrez Gene, and 153 broadly conserved miRNAs were used for the analysis.

3.2.2 The sampling procedure

By means of a randomization strategy, which had a p-value as output and is depicted in Figure 3.1, it was detected if a miRNA targeted a significantly higher number of genes in a gene list than expected by random expectation. For a gene list with n targets of a transcription factor (in our case REST) and a given miRNA miR-A, the number of miRNA-target gene pairs m_A was assessed from 72,770 TargetScanHuman miRNA-target gene pairs. In addition, the total number of miRNA-target gene pairs m_t was counted for the n genes and all 153 miRNAs. Afterwards, the same was done 10,000 times for n random genes to obtain a z_A and a z_t value for every run. Each time, a correcting factor r was calculated by the fraction of m_t and z_t and multiplied by z_A to obtain a bias-corrected quantity z_A* for comparison to m_A (see Section 3.4). We counted the number of times m_A was bigger or equal to z_A* and divided the value by 10,000 to generate the p-value of over-representation.

Since we performed the analysis described above for 153 miRNAs, we had to do multiple testing correction. We did this using the Benjamini and Hochberg method with one FDR for each miRNA as result (Benjamini and Hochbert, 1995).

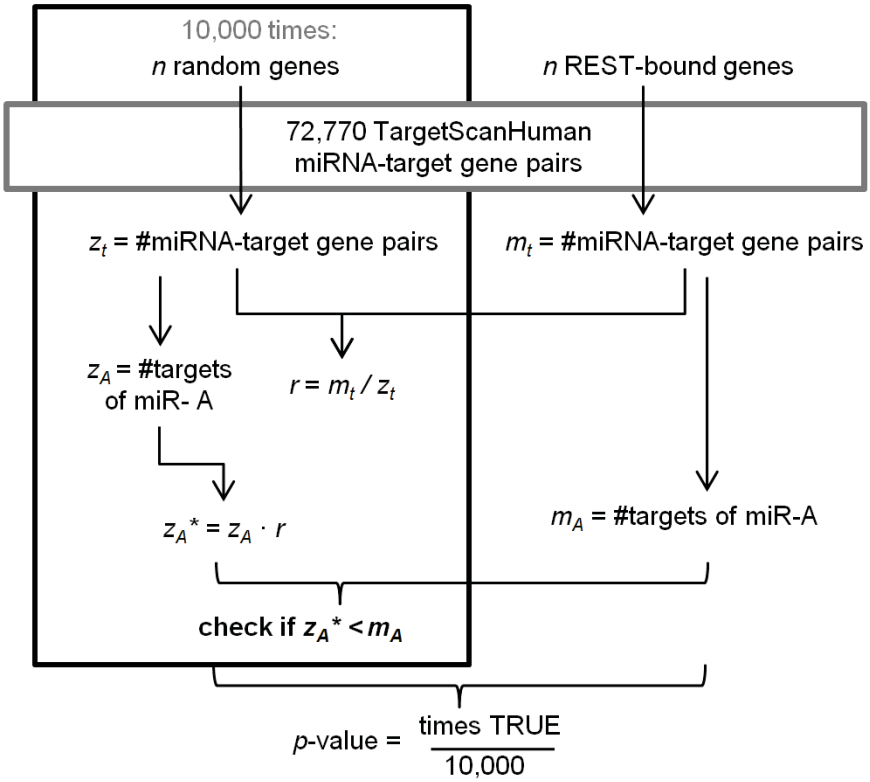

Figure 3.1: Calculation of over-represented miRNA target genes from gene lists (Gebhardt et al., 2014).

3.2.3 Implementing a general correction for 3'UTR biases in our algorithm

We wanted to illustrate the divergence between miRNA-target pair number of REST genes and the total background of the TargetScanHuman 6.2 dataset, and how it can be overcome. To do this for a test set with n genes, we performed the sampling 10,000 times as described in the previous section (i) without correcting z_A. We took the difference from m_A and the mean $z_{A,mean}$ of 10,000 z_A values. To make the result comparable to the other cell types, we calculated the percentage D_A from $z_{A,mean}$ (see Formula 3.1). It was plotted together with the D values from the other 152 miRNAs, as probability density curve yielding a Gaussian distribution.

$$D_A = \frac{(m_A - z_{A,mean}) \cdot 100}{z_{A,mean}} \tag{3.1}$$

(ii) we wanted to correct the 3'UTR length bias for the in (i) described procedure by

sampling 10,000 times from a classified background. The genes from the TargetScanHuman 6.2 dataset were sorted into 6 classes of 3'UTR length (0-700 bps, 700-1400 bps, 1400-2100 bps, 2100-3500 bps, 3500-8000 bps, > 8000 bps). The randomization was performed in a way that the 3'UTR length distribution of test set and random set were equal. This was achieved by counting the number of genes from each class in the test set and randomly drawing the same amount of genes from the respective background class. Afterwards, D_A was calculated and plotted as described above.

(iii) Finally, the sampling was performed exactly as described in Section 3.2.2, this time using the correcting factor and instead of building the mean of z_A, we calculated the average z_A* and used $z_{A,mean}*$ in Formula 3.1. The values of D were plotted for all 153 miRNAs.

Each described procedure (i-iii) was performed with the gene lists from eight cell types. Seven were taken from the ENCODE project (A549, ECC1, H1-hESC, HCT-116, HeLa-S3, HL-60, MCF-7) and the Jurkat T cells stemmed from Johnson et al. (2007).

3.2.4 Analyses on miRNA target predictions

Enrichment of REST-bound genes in predicted miRNA targets

We counted the number of miRNA targets n and the size of the REST gene subset m (according to the 15 ChIP-seq experiments). Afterwards, n random genes were taken from the TargetScanHuman 6.2 gene list and the number of REST targets z among them was counted. The procedure was repeated 10,000 times and we assessed how many times z was bigger than or equal to m. The resulting p-value was corrected for multiple testing using the Bonferroni method (Abdi, 2007).

Significance of filtering miRNA target predictions

The 'filtered' gene set was the union of all genes predicted to be regulated by REST and a miRNA enriched in REST-regulated targets. A list of experimentally validated miRNA target genes was downloaded from TarBase 6.0 (Vergoulis et al., 2012), a database with 38,384 miRNA-target pairs, 3,077 of them in TargetScanHuman 6.2. We compared the proportion of validated miRNA-target pairs from the total TargetScanHuman set and the filtered set, meaning the REST subset, calculated fold enrichment, and used Fisher's exact test to generate a corresponding p-value. This was done for each enrichment miRNA separately and for the union of all enrichment miRNAs with a minimum of 10 validated interactions in TarBase 6.0.

In addition, we made use of information from the binding sites of miRNAs from TargetScanHuman 6.2 to search for 'dual sites'. These are miRNA binding sites that appear in close proximity to each other (8 to 40 bps) and that often co-operate in regulation according to Grimson et al. (2007). Since the average miRNA binding site density and 3'UTR length is enhanced in REST-regulated genes (Section 3.3.2 and 3.3.3), we needed to control these parameters. First, we classified all genes according to their number of binding sites on the 3'UTR in range of class $i = 2, ..., 21$, omitting classes with higher numbers, which contained many transcripts with very long 3'UTRs. Second, we tested if the 3'UTR length distribution for the filtered set for each class had been shifted significantly towards longer 3'UTRs in comparison to the background by Wilcoxon rank-sum test. As expected, we found that in some classes there was a significant difference in 3'UTR length distribution. To be absolutely sure that the number of 'dual sites' from the filtered set was not improved by an enhanced miRNA binding site density, we used only classes for the following experiment that had p-values (from the Wilcoxon test) larger than 0.5, which were the classes $i = 2$, 3 and 13. For these classes we can expect that the miRNA binding site density is in the background set on average equal or larger than in the filtered set. We randomly selected equal numbers m_i of genes from the three classes for both sets (one-fifth of the genes in each class of the filtered set, 50, 42 and 31 for class $i = 2$, 3 and 13, respectively). In total $n = 123$ genes were selected from the filtered set and the background, respectively, and the total number S of 'dual sites' for each gene list was counted. The procedure was repeated 1,000 times yielding 1,000 values of S for both, test set and background.

3.2.5 Gaining insight into miRNA function

Small RNA-seq in various cell types

For nine of the 15 above mentioned cell types there were small RNA-sequencing (RNA-seq) data available in bedRnaElements-format, which contains a value for expression level and for statistical significance as well as the corresponding genomic positions of the called peaks. The peaks were assigned to miRNAs from miRBase (v.20, Kozomara and Griffiths-Jones (2011)). Afterwards, the expression values were normalized using the normalize.quantiles() function from *preprocessCore* library and log-transformed in R statistical programming language (R Development Core Team). The means of expression values over all members of a miRNA family were scaled by scale() and plotted with heatmap.2() from the *gplots* library.

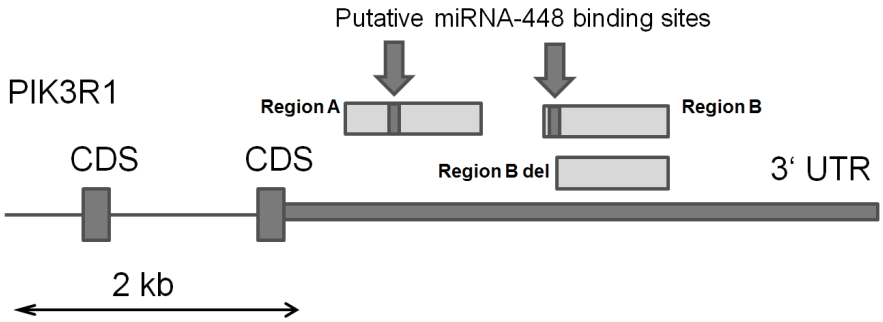

Figure 3.2: Dual reporter plasmids were created with Region A and Region B in the 3'UTR. Additionally there was a construct with a miR-448 binding site deleted version of Region B. The plasmid contained the reporter genes renilla under regulation of the PIK3R1 3'UTR and firefly. The shown CDS sequence parts are the coding regions of the PIK3R1 gene. Putative miR-448 binding sites are marked with big arrows and dark boxes. Figure adapted from (Gebhardt et al., 2014).

Clone counts in various cell types

The Atlas on Mammalian miRNA expression of Landgraf et al. (2007) was obtained by small RNA library sequencing and contains normalized clone counts for more than 150 different cell lines and cell types from tissues all over the human body. Only miRNAs with at least 10 copies detected over all tissues and present in TargetScanHuman 6.2 were considered. If the relative cloning frequency in a cell type was higher than 3% of the total clone count of the respective miRNA, it was designated as 'detected' there.

miRNA-UTR-assay: cloning, transfection and stimulation

The *in vitro* impact of miR-448 on PIK3R1 expression was examined by means of a miRNA-UTR-assay. 3'UTR regions of the PIK3R1 transcript were cloned with the following primers with restriction site overhangs (Figure 3.2):
Region A - chr5: 67593865-67593884; Region-A-forward: AGTActcgagGCCTGGTTTAGCCTGGATGT, RegionA-reverse GATgcggccgcCCCACCACCCCACTTGATAC
Region B - chr5: 67595300-67595319; Region-B-forward: GTCTctcgagTAGGGCAGGAGTGAGAGGTC, RegionB-reverse: TGAgcggccgcAAAACGACAAATGCGGTGGG

Region B was shortened to obtain a deletion of the putative miR-448 binding site using a PflI/XhoI digest. Afterwards, Klenow blunt end filling and re-ligation was performed. Genomic positions refer to human assembly hg19 of NCBI37 (February 2009).

The described fragments were cloned into a multiple cloning site of the dual reporter

plasmid. The applied plasmid, a modified version of the psiCHECK2 vector (Promega), contained the luciferase genes renilla under the control of the manipulated 3'UTR and firefly, both regulated by a constitutive promoter. The latter expression was used for normalization.

The reporter experiments were performed in HEK293 cells (n=6), seeded in 6-well plates in Dulbecco's modified Eagle's medium (DMEM, 106 cells per well). After transfection with the reporter plasmids (1 µg DNA/per well) and 3 µl Roti®-Fect (Carl Roth, P001.4) per well at 80% confluence, the cells were incubated for 24 hours. Then they were washed with equilibrated PBS, trypsinized (4 min at 37°C) and removed from the plates. After an additional step of washing and centrifuging the cells were re-suspended in DMEM medium and seeded in 96-well plates. These contained per well:
1. 10 pmol miR-448 (Invitrogen, hsa-miR-448, Assay-ID: MC10520),
2. 0.3 µl Lipofectamine® RNAiMAX (Life Technologies, Cat. 13778030)
3. 18 µl Opti-MEM® Medium (Life Technologies, Cat. 11058021).
For controls the same amount of Lipofectamine® RNAiMAX and Opti-MEM® Medium were used.

24 hours after transfection the cells were washed by removing the supernatant and adding 20 µl passive lysis buffer (Promega, Cat. E1941). The cells were incubated for 20 min at 20°C and the reporter assay was conducted as described by Hampf and Gossen (2006) with 100 µl Firefly buffer (Tricine 20 mM, $MgSO_4$ 2.67 mM, EDTA 100 µM, ATP 530 µM, DTT 33.3 mM, Coenzyme A 270 µM, D-Luciferin 470 µM, pH 7.8) and 100 µl Renilla buffer (NaCl 1.1 M, K_2HPO_4 220 mM, Na-EDTA 2.2 mM, BSA 6.58 mM, coelenterazine 1.43 µM, pH 5.1).

A Luminoskan luminometer (LabSystems) was used to measure emitted light after automated injection of the buffers with an in-house developed remote control for the Luminoskan luminometer. Relative light units of the renilla/firefly were re-scaled so that the mean of control of Region B equaled 1.0. A t-test was performed to assess significant differences in the measured values.

3.2.6 Extension of the approach

Integrating expression data and motif search

An unrefined REST target list in mouse and a gene list with genes additionally up-regulated upon REST knock-out were already generated in the course of the benchmarking procedure from Chapter 2.

Searching for motifs within REST peak regions of the ChIP-seq experiment of Arnold et al. (2012) the sequences of the successfully mapped regions were downloaded as FASTA-files from the UCSC Genome Browser (Kent et al., 2002) and analyzed with the HOMER findMotifs.pl functionality with a previously determined library of binding motifs (Heinz et al., 2010). Only genes with RE1 motifs in the corresponding peaks were considered for the motif-filtered list. The three lists were analyzed using the above implemented sampling procedure for search of miRNAs with enrichment in REST-regulated targets (Section 3.2.2).

Integrating DHS sites

The ENCODE project provides DNase I hypersensitive (DHS) site profiles for eight of the 15 cell lines with ChIP-seq data on REST binding. They were downloaded from the ENCODE repository (Suppl. Table S1) in BED-format. Overlapping regions from the two replicates of the REST ChIP-seq experiment were identified as described in Suppl. Methods *General Methods* yielding the first list of peaks. In the same way, we searched for an overlap of these peak regions with the DHS regions for our second peak list. Moreover, DHS region sequences were captured by means of the web platform *Galaxy* using the function 'Extract Genomic DNA'. Peaks with a RE1 site were extracted from the FASTA-formated result file using the motifScores() function from the *PWMEnrich* library (Stojnic and Diez, 2014) in R statistical programming language with parameters: cutoff = $log2(e^7)$, raw.scores=T and a positional weight matrix from the *PWMEnrich.Hsapiens.background* package (id = 'REST'). The peaks from DHS regions with RE1 motif yielded our third list. From the three resulting BED-formated files, peak-gene association was performed using the ranked method for consistency across the experiments. The gene lists were analyzed using the above implemented procedure for search of enrichment miRNAs (Section 3.2.2).

3.2.7 Setup of the web application

The search for miRNAs with enrichment in REST-regulated target genes was winded into a Python based web framework called Django (Django Version 1.5, 2013). Django allows to check the input for wrong characters and to perform tests on uploaded files to control size and format of the input. The peak-gene association interface forwards data in the required input format to the R functions, which perform the target gene and miRNA calling.

User provided gene lists are compared to a MySQL database that contains Entrez

IDs, Ensemble IDs, Gene Symbols and RefSeq gene Ids. If necessary, the identifiers are converted to Entrez IDs. In general, only Entrez IDs listed in the TargetScan data are forwarded.

Processed gene lists are passed to the main application, which performs the analysis for miRNAs with enriched targets in the respective gene list exactly as described in Section 3.2.2. It is coded in Perl and makes use of the CPAN modules 'List::Util', 'POSIX', 'SORT::Rank' and 'GD::Graph::histogram'. All output files are displayed by Django and are made available for download. The source code of the whole web application can be found in Suppl. Directory 1.

3.3 Characteristics of the underlying data

3.3.1 Properties of REST targets assessed from the ChIP-seq data

Results

The 15 generated lists of REST targets give a nice overview of the binding activity of REST in different cell lines. Table 3.1 shows that the number of REST target genes varied in the cell types. In the lung cell carcinoma cell line A549 8,356 genes with REST binding close by were detected, while in the HCT-116 and U87 cell lines only one-tenth of this number was achieved. 12,344 of 22,018 searched genes had a REST peak in proximity in any condition, which is a fraction of 56%.

We wanted to find out how specific the REST binding profile is to each cell type. The similarity of the 15 gene lists can be expressed by means of the Jaccard-index (see Figure 3.3), calculated as described in Suppl. Methods *General Methods*. According to the Jaccard-index, the cell types arranged in two clusters. The first cluster (Cl.I) contains cell types of non-neural origin with exception of the glioblastoma cell line U87. The second cluster (Cl.II) turns out to be dominated by neural cell types but also comprises cell line A549 and the leukemia cell line K562. It has to be emphasized that inside Cl.II the H1-neurons clustered away from the other cell lines.

Before proceeding with the analysis we defined a target classification similar to Bruce et al. (2009), which originated from ChIP-chip experiments on REST occupancy in eight cell types. They classified REST target genes to be detectable in all ('common'), in some ('restricted') or in only one ('unique') cell type. We wanted to assess if the genes from the Cl.I cell types were contained in the target sets of Cl.II and to what extent genes classified as 'unique' had impact on the clustering obtained with the Jaccard-

Table 3.1: Description of the cell lines used for the analyses and the corresponding numbers of identified possible target genes and size of TargetScanHuman 6.2 subsets.

Cell type	Nr. of REST target genes	Nr. of genes in TargetScanHuman 6.2	Description
A549^c	8356	5445	lung cell carcinoma
ECC1^c	1860	1260	endometrium adenocarcinoma
GM12878^n	1928	1264	B-lymphocyte
H1-hESC^n	2919	1892	embryonic stem cell
H1-neurons^n	3596	2656	neurons from H1-hESC
HCT-116^c	829	580	colorectal carcinoma
HeLa-S3^c	2091	1429	cervical carcinoma
HepG2^c	2639	1721	hepatocellular carcinoma
HL-60^c	1258	867	promyelocytic leukemia
K562^c	4230	2858	chronic myelogenous leukemia
MCF-7^c	1330	933	mammary gland, adenocarcinoma
PANC-1^c	1776	1179	pancreatic carcinoma
PFSK-1^c	4629	3103	cerebral brain tumor
SK-N-SH^c	6734	4516	neuroblastoma
U87^c	820	564	glioblastoma

c Karyotype cancer
n Karyotype normal

index. Figure 3.4 displays the fraction of genes shared by each pair of cell lines. The most striking observation is that H1-neurons shared fewest targets with the other cell types, although cell lines of neural origin were among them (Figure 3.4, compare H1-neurons from the y-axis with the cell types from the x-axis). Only cell line A549 and neuroblastoma cell line SK-N-SH had an appreciable overlap with H1-neurons and these were the two cell lines with the highest number of targets in the gene lists (see Table 3.1 and Figure 3.5A).

The second important observation from Figure 3.4 is that the gene lists from cluster Cl.I (refer to Cl.I on the y-axis and compare to all cell types on the x-axis) represented a subset of the targets from Cl.II to a certain extent, which can be deduced from the high fraction of common genes (blue color). In addition, we found that the number of 'unique' targets is quite proportional to the set size (see Figure 3.5A).

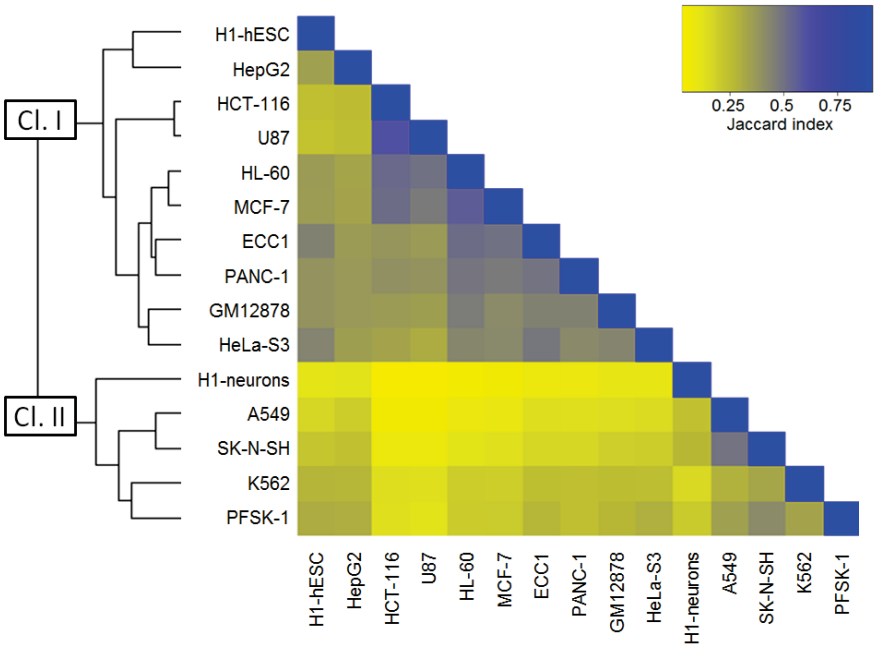

Figure 3.3: Potential REST target genes were compared using the Jaccard-index. Blue means that the list of genes is very similar between two cell types (Cl.= cluster).

221 targets were common to all 15 cell types. Since we found that the targets of H1-neurons are different from the other cell types, we also added a new class representing the number of targets common to all cell lines, excluding the H1-neurons. This class accounted for another 400 target genes. This is a strikingly high number when one considers that genes shared by 11, 12 and 13 cell types only sum up to 351 targets (see Suppl. Figure S3). In cluster Cl.I many cell lines did not have a single 'unique' gene. Glioblastoma cell line U87 and the colorectal carcinoma cell line HCT-116 even have more 'common' than 'restricted' targets.

The 'unique' gene sets were analyzed for enrichment in Gene Ontology terms in respect to all 12,344 REST targets. Cell line SK-N-SH and K562 showed significant enrichment in the terms 'regulation of transcription' (FDR $= 5.4 \cdot 10^{-5}$) and 'intracellular transport' (FDR $= 0.032$), respectively. Hepatocellular tumor cell line HepG2 was enriched in 'response to wounding' (FDR $= 2.2 \cdot 10^{-8}$) and A549 was shown to be involved in mitosis via the terms 'M phase of mitotic cell cycle' (FDR $= 1.1 \cdot 10^{-5}$) and 'nuclear division' (FDR $= 1.7 \cdot 10^{-5}$). The Gene Ontology terms designate possible biological processes that are specifically influenced by REST in the particular cell types. In cell types with

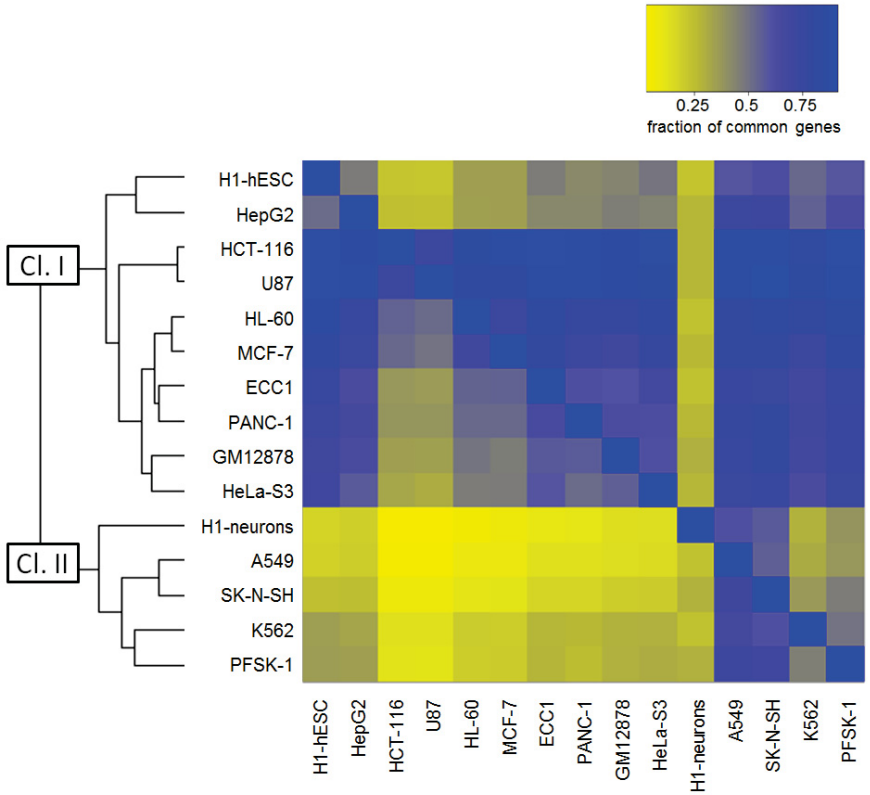

Figure 3.4: The gene lists of pairs of cell lines were compared and the fraction of genes in both cell types was calculated in respect to cell lines on the y-axis (horizontal values) and in respect to cell lines on the x-axis (vertical values). The targets of H1-neurons (y-axis) are included to a limited extent in gene lists of cell lines from Cl.I (horizontal, yellow values), but the genes from Cl.I gene lists (y-axis) are targets in most of the cell lines (horizontal, blue values with exception for H1-neurons (x-axis)).

very few 'unique' targets REST mostly performs the already known regulatory functions in repression of neural genes and intrinsic cellular processes.

Since REST is famous for its function as neural repressor, we made the naive assumption that the cell lines from the non-neural cluster Cl.I could exhibit a higher number of neural target genes than the cell lines of Cl.II, which contains the H1-neurons. A list of 456 brain-specific genes defined by Fang et al. (2009) was used to calculate significance of enrichment and fractions of neural genes in the 15 gene lists. The result was plotted together with the number of genes in each list to make the impact of the set size visible (Figure 3.5A, B and C).

All gene lists were enriched in neural targets. Enrichment p-values between 0.001

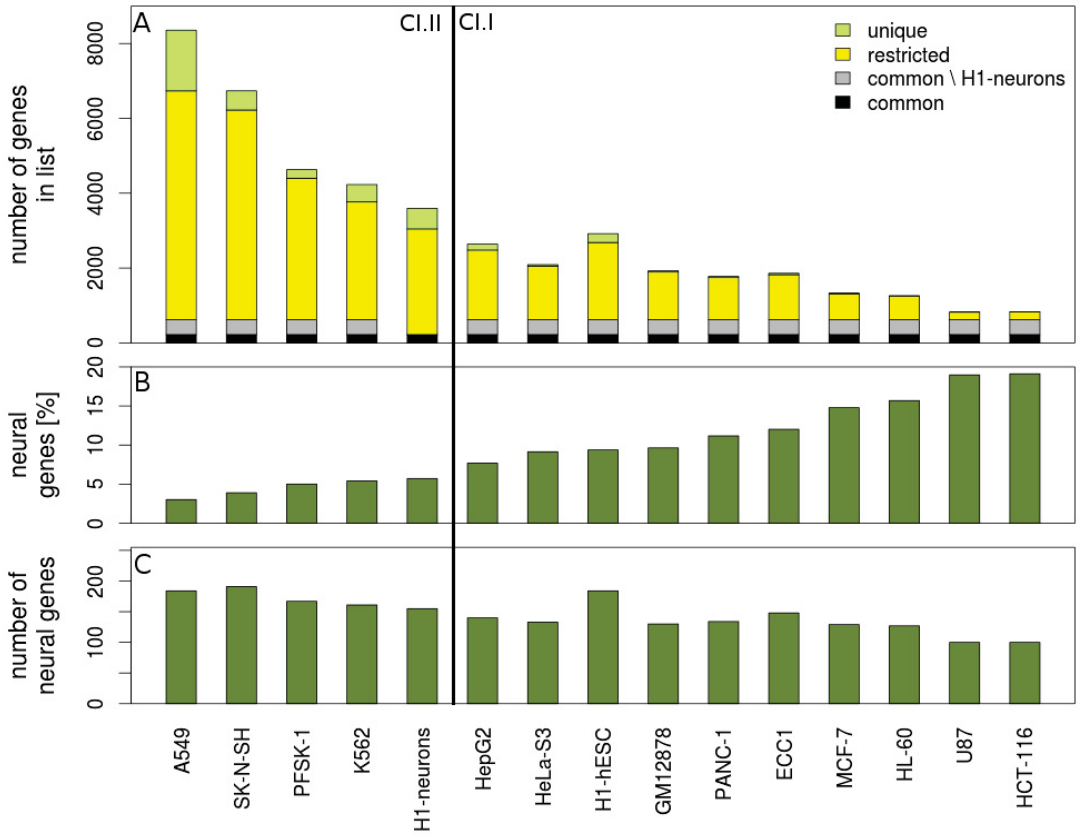

Figure 3.5: Composition of REST target lists from 15 cell types. Cl. II is the neural cluster, Cl. I is the cluster for non-neural cell types.
A) Total number of genes for every cell type and their classification into unique, restricted and common target genes according to Bruce et al. (2009). A fourth class are genes common to all cell types except H1-neurons.
B) Percentage of neural genes in respect to total set size.
C) Number of neural genes.
(Cl.= cluster).

for the A549 cell line and $2.48 \cdot 10^{-49}$ for cell line HCT-116 were achieved. Indeed the fraction of neural genes was higher in Cl.I than in cluster Cl.II (Figure 3.5B). The total number of neural genes, however, remained almost constant and the higher enrichment in Cl.I stems from the lower set size. It is noteworthy, that the 221 common targets, which are shared among all cell lines, had a higher enrichment in neural genes with 24.3% (p-value $= 5.18 \cdot 10^{-21}$) than the gene lists from each single cell line. The neural repressor function of REST can be found here.

Discussion

The high number of 12,344 REST-bound genes for the union of all 15 lists is in accordance with a former study, which extrapolated the number of REST binding sites from about 1% of the genome and estimated a value of around 25,000. We confirm the observation that the REST target lists are composed of (i) a core of 'common' genes, which in our case form about 1.8% of all targets, (ii) a number of 'unique' genes, which is proportional to the set size, and (iii) a majority of 'restricted' genes. Bruce et al. (2009) identified about 10% of the REST targets as 'common' genes, but the higher number was to be expected because they did a comparison of only 8 cell lines.

The ChIP-seq experiments monitored a high variability in the number of targets in different cell types. Bruce et al. (2009) uncovered the molecular basis for the relationship of RE1 motif and REST occupancy. They found that 'common' REST binding sites tend to feature a canonical binding motif with a higher binding affinity resulting in higher REST occupancy values when compared to 'restricted' and 'unique' binding sites, which usually have more atypical binding motifs (see Section 1.2.2). Thus, we assume that in cell types with many targets REST has a much higher concentration, or altered binding specificity by co-factors in comparison to cell types with few targets. Otherwise it would not be capable of binding to the high amount of non-canonical binding motifs. This matches the observation that small target lists form subsets of larger lists.

According to our observations, the target set in H1-neurons differs a lot from the REST-bound genes in other cell types. It is even very different from other neural cell types such as U87, SK-N-SH and PFSK-1. All of these cell types have a certain need of neural gene expression. Thus, neural gene expression in H1-neurons cannot be the only explanation for the diverging target set. Another possible reason is the fact that the H1-neurons are a primary culture while all other cell types in the experiment are established cell lines. The proteomes of primary cultures and cell lines differ in respect to metabolic pathways, cell cycle-associated functions, and cell type specific enzymes (Pan et al., 2009).

3.3.2 3'UTR length bias in REST target genes

Results

The likelihood of finding a certain miRNA binding site on a 3'UTR is generally higher if the 3'UTR is longer. The TargetScan algorithm compensates this bias by giving more weight to predicted binding sites that are not in the center of the 3'UTR. Nevertheless, since the 3'UTR length could have significant impact on the final results in the enrichment analysis, we took a closer look at the 3'UTR length distribution of REST targets in comparison to all genes (Figure 3.6).

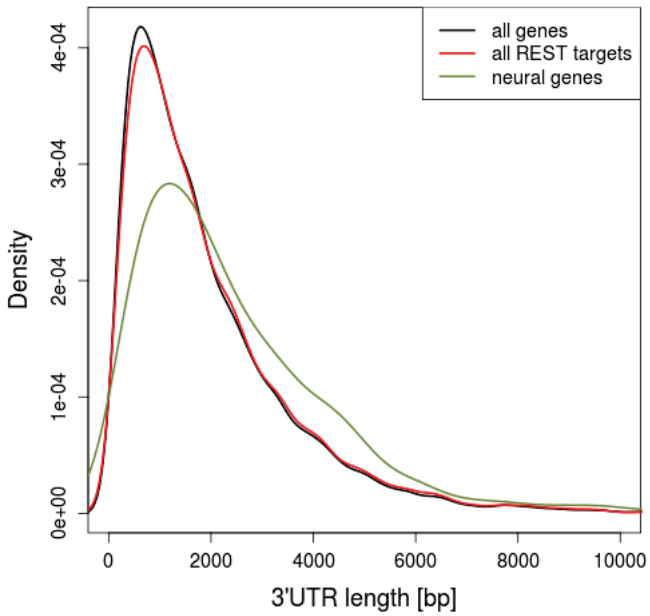

Figure 3.6: Probability density curves derived from the 3'UTR length distributions of all genes, the REST targets and a neural subset from TargetScanHuman 6.2. The curves of the neural genes and the REST target genes are shifted to the right.

The 3'UTR length distribution of REST target genes was shifted to the right in comparison to this of all genes in the TargetScanHuman 6.2 dataset. Due to the fact that the REST targets comprise about 70% of the dataset, the shift was very modest. According to the one-sided Wilcoxon rank-sum test, it was significant with a p-value of 0.0068. The difference became more obvious when the test was done without restriction to genes with miRNA binding site predictions (p-value $= 4.30 \cdot 10^{-26}$). Neural genes (as defined in Fang et al., 2009) had a length distribution that was clearly shifted towards long 3'UTRs (Figure 3.6).

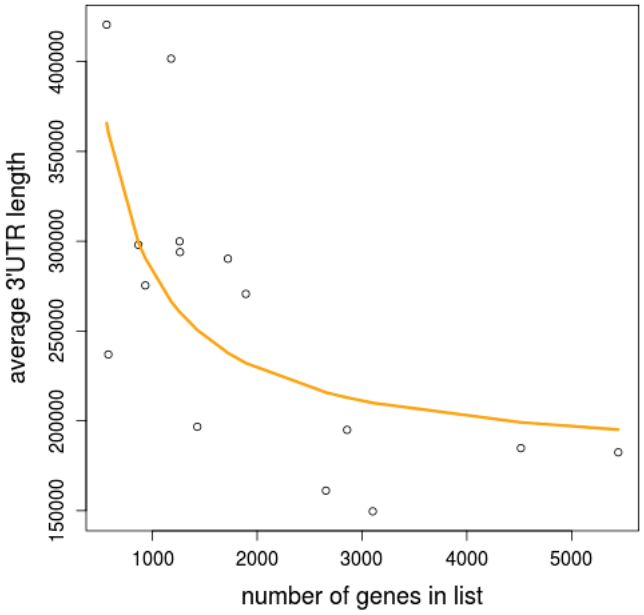

Figure 3.7: Average 3'UTR length of the genes on 15 REST target lists is plotted as a function of the number of genes in the respective list. The orange curve is a non-linear fit $f(x) = I(1/x \cdot a) + b$, convergence tolerance $= 7.46 \cdot 10^{-7}$, a = 107,434,935, b = 175,361. There is a negative correlation of set size and average 3'UTR length.

We plotted the average 3'UTR length for the 15 cell types separately against the gene list size. The relation is depicted in Figure 3.7 and without being completely accurate it shows that the 3'UTR length bias was indeed more distinct in cell types with small REST target lists.

Discussion

According to the analysis, REST targets have longer 3'UTRs on average than the background of all human genes. One reason is that the target lists exhibit high numbers of neural genes, which have very long 3'UTRs on average (see Figure 3.6). It has been shown that tissue-specific genes have longer 3'UTRs than genes that regulate basic cellular processes (Stark et al., 2005).

From Section 3.3.1, we know that the lists of REST target genes from the 15 cell types vary in set size and that smaller lists tend to be subsets of larger lists, containing higher fractions of neural genes. Thus, we expected to find a negative correlation of set size and average 3'UTR length, which we could demonstrate. A correction for this 3'UTR length bias had to be included in the algorithm as discussed in Section 3.4.

3.3.3 miRNA binding site density bias in REST target genes

Results

3'UTR length is only one parameter with impact on the number of predicted miRNA binding sites. We additionally observed a difference in miRNA binding site density between the 3'UTRs of REST target genes and background, which exists due to biological needs. E.g., subsets of genes could be under tight control of several miRNAs and, therefore, exhibit higher miRNA binding site densities in their 3'UTR than other genes. In our simulation approach we monitor miRNA-gene relations without considering the number of times one certain miRNA binds a 3'UTR. We focus on the number of genes with 3'UTRs bound by a specific miRNA instead of the number of miRNA binding sites in the 3'UTR. In Figure 3.8 we plotted the average number of predicted miRNAs targeting a 3'UTR for transcripts of REST-bound genes and the background, and split the genes into classes to reduce or possibly eliminate the impact of 3'UTR length on the result. Classes with 3'UTRs longer than 10,000 bps contained only few genes and are not shown. The average number of miRNAs per 3'UTR of REST targets was slightly higher in most classes than in the background.

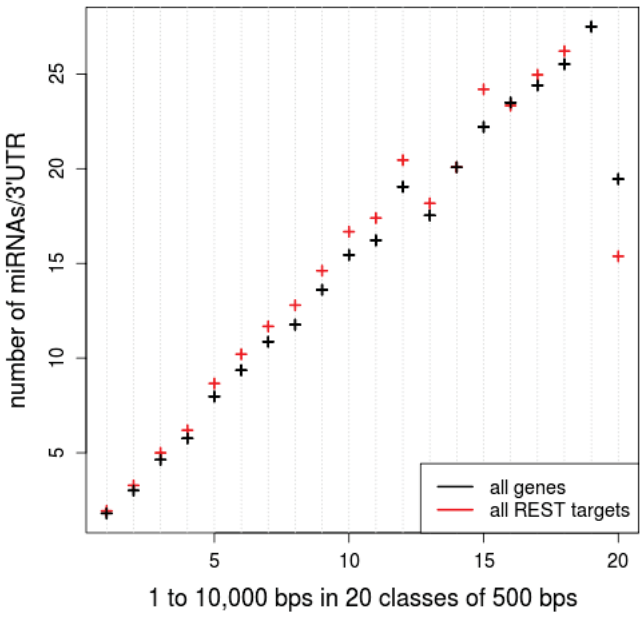

Figure 3.8: Average miRNA count per 3'UTR. The counts were compared for all genes and REST targets. To do this, they were split into 20 classes according to the lengths of their 3'UTRs from 1 to 10,000 bps in steps of 500 bps.

Discussion

We were able to show that in addition to longer 3'UTRs REST targets often have more regulating miRNAs predicted per 3'UTR than the average background. Neural genes have not only longer 3'UTRs on average but also more miRNA binding sites than genes with less specificity (Stark et al., 2005).

The goal is to design the analysis in a way that it can be applied to any gene list. A universal approach is needed to normalize 3'UTR length and miRNA binding site density. In the following section we address this issue.

3.4 Implementing a general correction for 3'UTR biases in our algorithm

Results

Initially, we approached the detection of over-represented miRNA target genes in gene lists using a simple simulation procedure. To estimate if the number of target genes m_A shared by REST and a certain miRNA miR-A was significantly large, this number m_A should be tested n times (n = 10,000) against a random set (z_A) of the same size. Random sampling should be performed on all 72,770 predicted miRNA-gene interactions of the TargetScanHuman 6.2 dataset. The result of the 10,000 comparisons (if $z_A < m_A$) was meant to be a p-value for over-representation (see Section 3.2.2).

We first ran this simulation without considering the above explained biases (Sections 3.3.2 and 3.3.3) to illustrate the divergence in miRNA binding site distribution. The number of genes m_A targeted per miRNA miR-A in n genes from the test set was counted. The same was done in the course of 10,000 randomizations for n genes from the total background and an average gene count was calculated ($z_{A,mean}$). The difference of m_A and $z_{A,mean}$ (as percentage D_A of m_A) was plotted for 153 miRNAs as density distribution (Figure 3.9A, see Section 3.2.3i for details).

We did this for seven of the 15 ENCODE cell types and one gene list derived from a ChIP-seq analysis in Jurkat T cells, coloring the curves according to the amount of genes in the lists. In Figure 3.9A dark gray curves from cell types with large target lists showed their maximum density closer to zero than light gray curves from smaller sets, but were still far away. The larger the size of the test set, the closer the behavior to the average background.

Sampling cannot be accurate as long as the distributions of miRNA-gene pairs in test and background set differ in properties relevant to the number of miRNA targets present

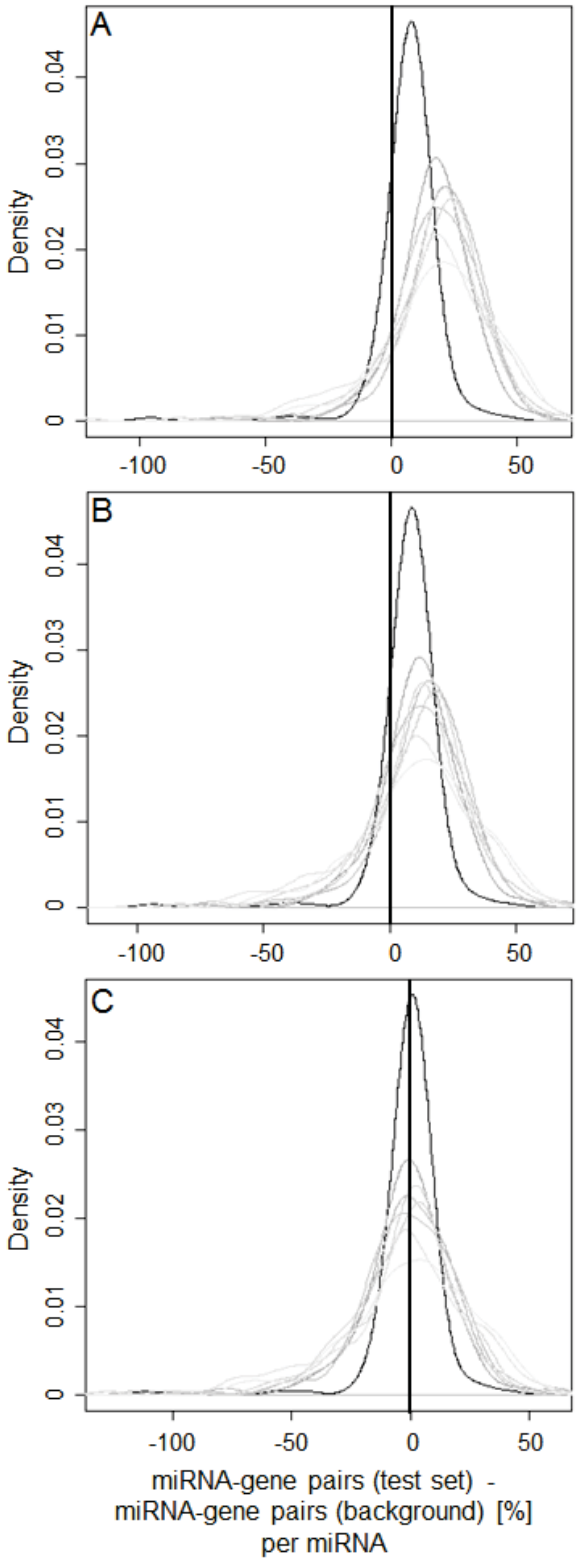

Figure 3.9: Correcting the bias in number of miRNA binding sites. Eight REST target gene lists were taken from the ENCODE project and (Johnson et al., 2007). For each gene list we calculated the difference of miRNA-gene pair numbers for REST targets and average background (measured in 10,000 randomizations) per miRNA as percentage of miRNA-gene pair count of the REST targets. Afterwards, we plotted the probability density curves of the values for all miRNAs. We did this for each cell line separately (see Section 3.2.3). The colors of the curves range from dark to light gray. The darker, the more genes were in the gene list.
A Not corrected.
B Corrected for 3'UTR length by sorting all genes into 6 length classes.
C Using the correcting factor. The number of genes with binding sites for a certain miRNA is corrected with the help of the total number of miRNA-gene pairs from test and background set (see Sections 3.2.2 and 3.2.3). this procedure achieves a centering of the maxima of the curves around zero.

in 3'UTRs, such as the length of the 3'UTR. It is obvious that the shift in miRNA-gene pair distribution was related to such a difference in 3'UTR length distribution. We tried to eliminate the bias by sampling the n random genes in six predefined 3'UTR length classes in a way that would make the 3'UTR length distribution of test and background set equal (see Section 3.2.3ii for details). Figure 3.9B shows that this procedure, although it moved the curves to the left, was not sufficient to eliminate the difference in miRNA-gene pair distribution of target sets and background.

To solve the problem, we applied a normalization approach that is explained in Section 3.2.2 and Figure 3.1. In short, it made use of the fraction r of total miRNA-gene pair count (sum of all 153 miRNAs) of the test set and of the random sets, to correct for the difference in miRNA-gene pair distribution (refer to Section 3.2.3iii for details about the calculation of the curves in Figure 3.9C). Figure 3.9C shows that this procedure erased the difference in miRNA-gene pair distribution of target gene set and background. The maxima of the curves appeared almost centered above zero.

Discussion

If our simulations are done without bias correction, this will result in an under-estimation of z_A, the number of targets of a certain miRNA in the random set, in each randomization step. As a consequence, all p-values will be smaller than they should be and more miRNAs will be claimed to have significantly over-represented targets in the respective gene list. Thus, it is absolutely essential to erase the difference in miRNA-gene pair distribution of test sets and random set.

There was a correlation of the target set size with the grade of shift that the curves in Figure 3.9A exhibit. As shown in Section 3.3.2 the 3'UTR length of REST genes is on average higher than that of the background genes and there is a correlation to the set size. As a result, it is to be expected that a great deal of the shift in the probability density curves is caused by differences in average 3'UTR length.

Sampling in 3'UTR classes helps to move the maximum of the curves closer to zero. This procedure corrects the 3'UTR length bias but not the bias in miRNA density. To obtain a good simulation result, the random sets must be chosen in a way that reproduces the length and miRNA density distribution of the test set. However, this is not possible due to the small size of the background.

The developers of the mirBridge method (see Section 1.3.3) found a way to correct for multiple biases, which they had identified in the analysis. They calculated the Euclidean distance of 3'UTR length, GC-content and conservation for all 3'UTRs in the dataset and sampled on genes with a close distance. Now we know that we would likewise need

to correct miRNA density. This would be possible, but the analysis is getting more complicated with each correction step.

The normalization approach from Figure 3.9C works well. An advantage of the normalization by means of the correcting factor r is that the difference in miRNA-gene pair distribution of test set and background will be corrected independently of its origin. Even if there were other biases than 3'UTR length and miRNA density that we have overseen so far, they would be corrected. The described procedure makes the algorithm very fast in comparison to other approaches. Shalgi et al. (2007) made use of a cumulative hyper-geometric distribution and a randomization approach (see Section 1.3.3) to find over-represented miRNA targets. Both approaches have the strong advantage that they circumvent the correction of the above mentioned biases. However, solving binomial coefficients and running the by Shalgi et al. (2007) described randomizations is very time consuming and needs a lot of computational power. In comparison our method is much faster and more concise.

The miTEA web application does miRNA target enrichment analysis, but it cannot be applied to the analysis of target lists from ChIP-seq data because they are not ranked. Hence, we do not need to compare it directly to our approach.

Limitations of the approach and alternatives

Here we will discuss the limitations of our approach and will compare it to other methods that study over-representation of miRNA targets in gene lists. We will also discuss variations of our approach and its possible shortcomings.

Regarding the miRNA target predictions, apart from the fact that many false positive predictions are included in the dataset, these predictions comprise only miRNA binding sites from 3'UTRs and not from other gene regions such as 5'UTR, CDS or introns.

Our method uses conserved miRNA binding site predictions from TargetScanHuman. As far as we know, no other methods makes full use of non-conserved miRNA binding site predictions of TargetScan. The mirBridge method (Tsang et al., 2010) is restricted to targets with at least one conserved seed-match site or one site with context score of at least 68, therefore, using only a small subset of non-conserved binding sites. Shalgi et al. (2007) utilized only conserved binding sites, just as we did.

In our case integrating non-conserved miRNA binding site predictions from TargetScanHuman into the analysis would slow down the method because for multiple testing correction the p-values generated are multiplied by factors in range of the number of searched miRNAs, which is about 1,500 for human, when the non-conserved data are considered. At least 100,000 randomizations would need to be done and only a very

significant over-representation would be detected. In most of the analyses no significant results would be obtained. Given these reasons and given the low quality of the non-conserved predictions, we decided against using such data.

Looking for enrichment of miRNA binding sites in the 3'UTRs of a gene set instead of the number of genes with at least one miRNA binding site would be possible, but since the impact of miRNAs targeting one gene multiple times cannot be estimated from the resulting p-values we preferred looking for over-representation of miRNA target genes.

The size of the input gene list should not be too small to obtain a good approximation of the regulatory impact of the transcription factor. If it is very small, it should have a very high precision, but an input gene list should not be too large, either. As we saw above, due to the restricted size of the TargetScanHuman background, the properties of a very large gene list are very similar to the background. As a result, it is unlikely to find significant enrichment for large gene lists.

3.5 Detecting over-represented miRNAs in gene lists

Results

Now, having an algorithm at hand to search for enriched miRNA targets in REST gene lists (see Section 3.2.2), we calculated the significance of over-representation of miRNA targets in comparison to a random background with one FDR per miRNA as output. From the 153 miRNAs, only results with FDRs smaller than the arbitrarily chosen significance level of 0.1 are presented in Table 3.2. It comprises 20 miRNAs. A stricter cutoff of 0.01 would still have produced a table with five entries. miRNAs with over-represented targets in REST gene lists will be referred to as 'enrichment miRNAs'.

No enrichment was found in cell line A549, therefore, it is not depicted in Table 3.2 and it is excluded from the following discussions.

Table 3.2 comprises two miRNA pairs that have overlapping miRNA seeds and one miRNA with an overlap to a miRNA that is not in the table (see Section 1.3.1). The seed overlap led to high numbers of common target genes, depicted in Figure 3.10. In the figure the overlapping seed region can be seen.

Table 3.2: miRNAs with significantly over-represented targets in the genes of 14 REST target lists (FDR < 0.1). Cell types with origin in neural tissue are marked in gray. Table from (Gebhardt et al., 2014).

miRNA family	Times found	ECC1	GM12878	H1-hESC	H1-neurons	HCT-116	HeLa-S3	HepG2	HL-60	K562	MCF-7	PANC-1	PFSK-1	SK-N-SH	U87
miR-101/101ab	1	-	-	-	0.058	-	-	-	-	-	-	-	-	-	-
miR-129-5p/129ab-5p	8	-	0.008	0.054	0.015	-	0.031	0.031	-	0.010	-	0.092	-	-	0.015
miR-132/212/212-3p	1	-	-	-	0.078	-	-	-	-	-	-	-	-	-	-
miR-138/138ab	6	-	-	0.070	-	0.096	-	0.097	-	0.010	-	0.097	-	-	0.015
miR-139-5p	1	-	-	-	0.022	-	-	-	-	-	-	-	-	-	-
miR-153*	12	0.077	0.008	0.005	0.010	0.008	0.094	0.036	-	0.015	0.008	0.005	-	0.015	0.015
miR-185 family	8	0.061	0.058	0.096	-	0.005	0.008	-	-	-	0.004	0.005	-	-	0.048
miR-190/190ab	1	-	-	0.079	-	-	-	-	-	-	-	-	-	-	-
miR-208 family**	1	-	-	-	0.023	-	-	-	-	-	-	-	-	-	-
miR-217	1	-	-	-	0.010	-	-	-	-	-	-	-	-	-	-
miR-218/218a	10	0.054	0.005	0.005	-	0.005	0.073	-	0.069	-	0.004	0.086	0.065	-	0.008
miR-300/381/539-3p	2	-	-	-	0.019	-	-	-	-	-	-	-	-	0.082	-
miR-326/330/330-5p	1	-	-	-	-	-	-	-	-	-	-	-	-	-	0.048
miR-329 family	4	-	-	0.079	-	0.052	-	-	-	-	0.021	0.092	-	-	-
miR-34 family	1	-	-	-	-	-	-	-	-	-	-	-	-	-	0.086
miR-374ab	1	-	-	-	-	-	-	0.061	-	-	-	-	-	-	-
miR-421	2	-	0.071	-	-	0.064	-	-	-	-	-	-	-	-	-
miR-448/448-3p*	13	0.015	0.008	0.005	0.022	0.005	0.008	0.015	0.015	0.019	0.008	0.005	-	0.015	0.015
miR-499-5p**	1	-	-	-	0.018	-	-	-	-	-	-	-	-	-	-
miR-543	2	-	-	-	0.064	-	-	-	-	-	-	-	0.082	-	-

No significant enrichment miRNAs were detected for A549. This cell line was excluded from the table.
∗,∗∗ These miRNAs have overlapping seeds (non independent results, see this Section and Section 1.3.1)

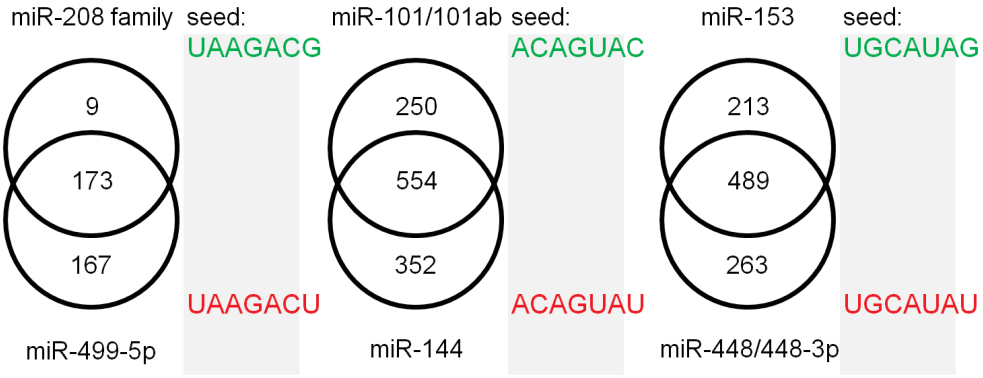

Figure 3.10: Pairs of miRNAs with overlapping seed sequences are presented with their numbers of target genes and the number of genes that are shared. The seed sequence can be found next to the Venn diagrams. The gray shaded area illustrates the overlapping seed region.

As a test, we performed an alternative search for enrichment miRNAs excluding target genes with overlapping predicted binding sites. Although this reduced the set size of the gene lists enormously, two of the detected miRNAs, miR-101 and miR-448, still had enriched targets in the REST gene list with a FDR < 0.2 (miR-101 in H1-neurons FDR $= 0.134$, for miR-448 see Suppl. Table S6). For the other miRNAs no target over-representation could be detected. This test ensured that our results are not simply due to overlapping seeds.

We also examined the converse enrichment of REST-bound genes in the lists of miRNA targets for each of the 20 enrichment miRNAs. According to the ChIP-seq data, 11 out of 20 enrichment miRNAs had an over-representation of REST-bound genes in their predicted target genes from TargetScanHuman 6.2 (Bonferroni adjusted p-value < 0.05, see Methods Section 3.2.4 *Enrichment of REST-bound genes in predicted miRNA targets*, Suppl. File 2).

From 20 miRNAs that had enriched target genes with FDRs above the significance level 10 appear at least twice. miR-129, miR-138, miR-153, miR-185, miR-218 and miR-448 had over-represented targets 8, 6, 12, 8, 10 and 13 times in 14 cell types, respectively. One could argue that these high numbers of common enrichment miRNAs must originate from target genes of REST that are shared among the cell types. To assess if this is true, we calculated the Jaccard-index based on the set of enrichment miRNAs for each cell type and plotted how the cell types cluster according to the indices (Figure 3.11) as

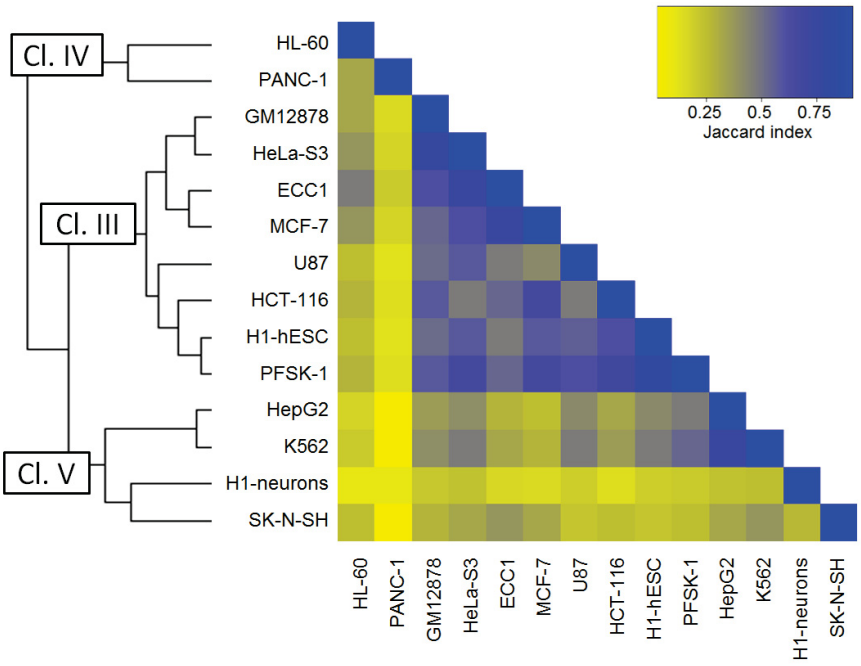

Figure 3.11: In targets of 14 REST gene lists enriched miRNAs were compared using the Jaccard-index. Blue means list of miRNAs is very similar between two cell types (Cl.= cluster).

we have done for the genes in Figure 3.3. Instead of separating into a neural and a non-neural cluster as in the case of Jaccard-indices from genes, in Figure 3.11 three clusters can be found. Cl.III comprises eight cell types of mixed origin. Cl.IV clusters far away from this and Cl.V contains a neural sub-cluster with H1-neurons and the SK-N-SH neuroblastoma cell line. Regarding the H1-neurons, this result was to be expected because the enrichment miRNA profile of the H1-neurons in Table 3.2 differs a lot from the other cell types. For example, over-representation in miR-101, miR-132, miR-139, miR-208, miR-217 and miR-449 targets was exclusive to H1-neurons. miR-300 and miR-543 were only shared by cell lines of neural origin. Enrichment miRNAs miR-138, miR-185 and miR-218 were not found in H1-neurons and other neural cell types except in U87. The glioblastoma cell U87 line had also clustered away from the neural cluster according to the Jaccard-index by genes exhibiting a non-neural profile (Section 3.3.1).

To be able to compare the gene and miRNA Jaccard-indices we plotted the Jaccard-index of the genes from Figure 3.3 against the Jaccard-indices calculated for the miRNAs in Figure 3.12.

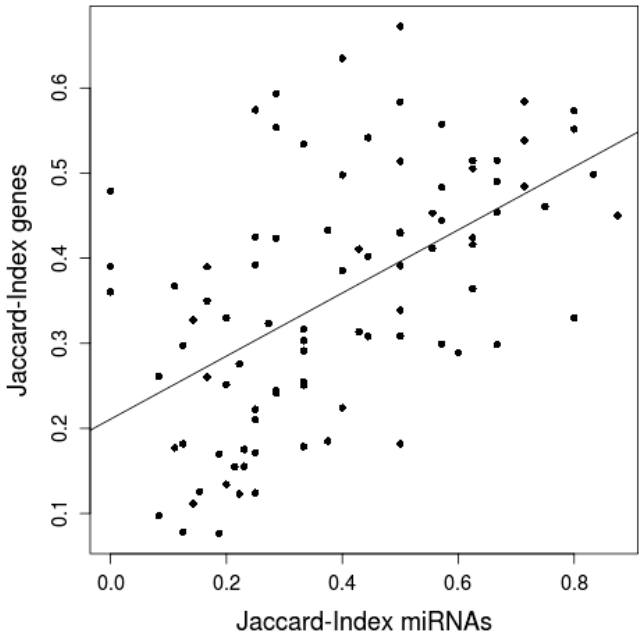

Figure 3.12: Jaccard-index of genes is plotted against the Jaccard-index of enrichment miRNAs for each possible pair of cell lines. Coefficients for the linear regression (black line): intercept = 0.21, slope = 0.37, $R^2 = 0.29$.

If the number of common miRNAs found for the cell types was solely dependent on the shared genes of the cell types, one would expect a clear linear relation between the two parameters. Some data points appear far away from the straight regression line and the regression coefficient of 0.29 confirms that no clear correlation can be found. To pick an example, the Jaccard-index of genes for HepG2 and K562 was only 0.33, pointing to a low number of shared targets among the cell types. The list of enrichment miRNAs in HepG2, however, comprised five miRNAs of which four could be found in K562, resulting in a Jaccard-index of 0.8 for enrichment miRNAs.

Discussion

Many enrichment miRNAs were detected in multiple cell types, and by means of the Jaccard-indices we showed that these findings cannot only be caused by common genes in the gene lists. We assume that REST and the enrichment miRNAs are part of a regulatory network around common modules and that, depending on the cell type, different subsets of regulators are active.

In former studies, researchers were able to identify miRNAs with targets over-represented

in genes potentially regulated by REST (Shalgi et al., 2007). With the help of computationally detected RE1 binding motifs in the promoter region of genes, a target gene list was generated. The RE1 motifs yield a genome-wide view on possible REST binding sites in a cell type independent manner. Without experimental evidence one will never know if the binding sites can be bound *in vivo*, but due to the long and well conserved RE1 motif that tends to be situated in promoter regions, the results are trustworthy. According to Shalgi et al. (2007), miR-153, miR-448 and miR-326 were significantly co-regulating REST target genes. We confirmed these results with our analysis. Results from Shalgi et al. (2007) also proposing miR-7, miR-133 and miR-135 could not be reproduced by us. The reason for the difference between our results and those from the study of Shalgi et al. (2007) can be due to the fact that we made use of experimental data to create our gene list. The detected REST binding sites build a cell type specific profile and include canonical and non-canonical sites. In conjunction with the study of Shalgi et al. (2007) we can conclude that REST and miR-153 or miR-448 share a significant amount of target genes on a genome-wide and cell type independent scale. If this is the case, it is very likely to detect them as enrichment miRNAs in almost any tested cell type, which is precisely what we found. Thus we demonstrated how the application of cell type specific experimental data such as ChIP-seq data can reveal relations that are cell type independent. This is quite important because for practical experiments a researcher will always have to decide on certain conditions even if he is interested in much more global relations. In other words, we propose that one does not need to test all possible cell types and conditions with ChIP-seq (or other methods to detect regulatory interactions) to gain insights into the full gene regulatory network.

In Table 3.2 we found five enrichment miRNAs whose miRNA seeds overlap with the seed of another miRNA. An over-representation of targets could be confirmed for miR-101 and miR-448 by elimination of targets common to both partners from the respective gene lists. For miR-153, miR-208 and miR-499 we cannot state with certainty, if they are true enrichment miRNAs. For further discussions, we will nevertheless act on the assumption that they are. Interestingly, these five miRNAs were detected in H1-neurons. It is possible that the overlap of the miRNA seeds has a biological function. However, this is only speculation that needs to be examined thoroughly.

That the set of enrichment miRNAs in H1-neurons is different from other cell types can be explained by the differing target gene lists (see Section 3.3.1). It was to be expected that other parts of the regulatory network need to be activated or repressed

to confine neural and non-neural processes. miRNAs, such as miR-217, are of special interest, because in this case over-representation is exclusive to H1-neurons. It is a predicted regulator of REST and it is known to be expressed in H1-neurons. Relations such as these will be examined in the following section.

3.6 Gaining insight into miRNA function

3.6.1 Enrichment miRNAs, their expression and REST regulation

Results

The search for enrichment miRNAs provides information about miRNA-target relations, but there are more aspects around the identified modules that can be analyzed. The functional relevance of the enrichment miRNAs can be interpreted by means of the miRNA expression pattern. The ENCODE project delivered associated small RNA sequencing data from nine of the 15 analyzed cell types, among them H1-neurons (see Section 3.2.5 *Small RNA-seq in various cell types*). Suppl. Figure S4, which gives an impression of the expression values of all 153 miRNAs, shows obvious differences between H1-neurons and the other cell types. Figure 3.13 concentrates on the expression of the 20 enrichment miRNAs in the respective cell types. Among these, elevated expression of miR-217, miR-448 and miR-153, and reduced expression of miR-101 and miR-329 contributed most to that difference. miR-421, miR-139 and miR-374 were highly expressed and miR-129, miR-208 and miR-449 were almost not expressed in all cell lines.

When we compare the miRNA expression to the results of over-representation in Table 3.2, we find that in most cases an enrichment miRNA was not detected in the cell type where it was highly expressed. Exceptions were e.g. miR-374 in HepG2, miR-185 in MCF-7 and miR-448 in H1-hESC. In H1-neurons the situation was different. Here at least four enrichment miRNAs miR-139, miR-153, miR-217 and miR-448 were expressed. Depending on the choice of threshold for 'expressed' miRNAs, we obtained a p-value between 0.54 and 0.11 for the over-representation of H1-neuron expressed miRNAs in the enrichment miRNA set (Fisher's exact test, see Suppl. Figure S5 for density distribution of expression values in H1-neurons). This turns out to be rather a tendency.

There are many more kinds of neural cell types. We performed a second test on over-representation of neural miRNAs in the 20 enrichment miRNAs by means of the Atlas on Mammalian miRNA Expression (see Section 3.2.5 *Clone counts in various cell types*, Landgraf et al., 2007). 16 of the 20 miRNAs were contained in this repository of miRNA

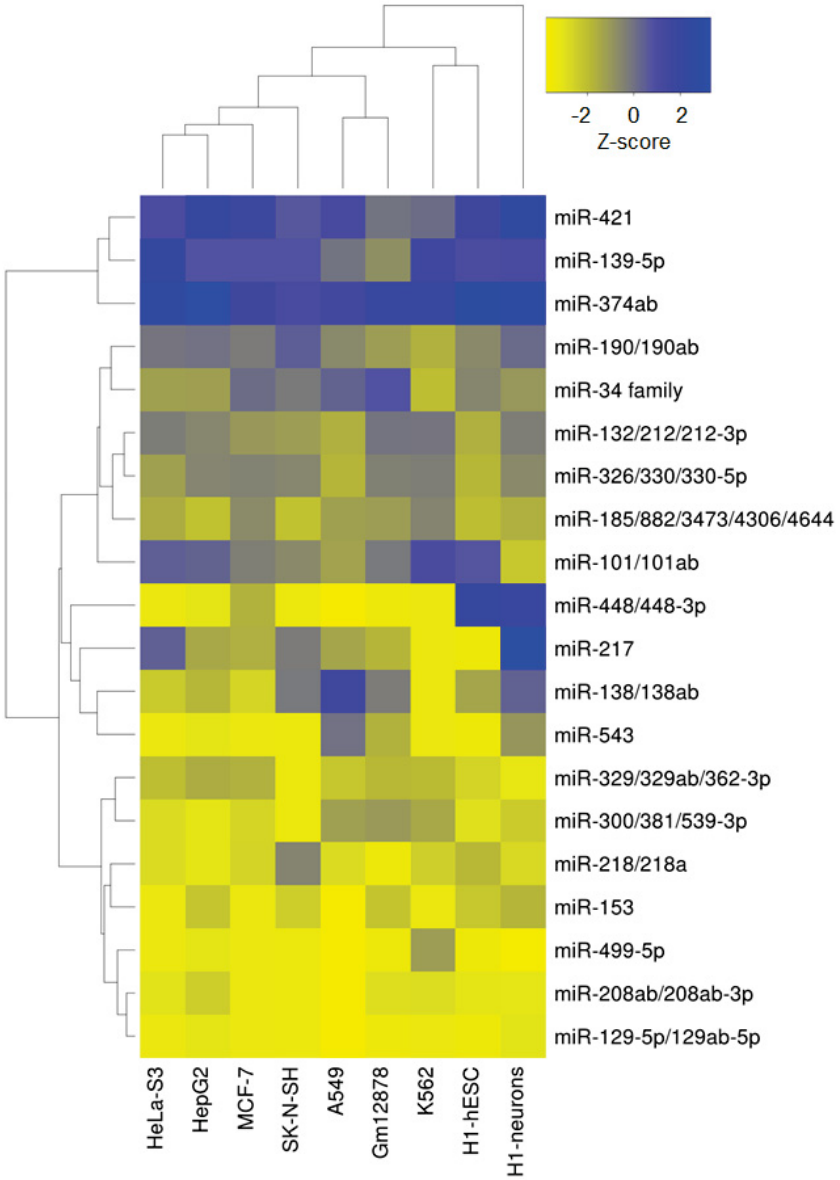

Figure 3.13: Heatmap and dendrogram of enrichment miRNA expression according to small RNA-seq on nine ENCODE cell types.

clone frequencies, of which 9 were identified as 'detected', some of them being even neural-specific according to the literature (Table 3.3). This corresponded to a p-value of 0.05. Using a newer repository for tissue specific miRNAs 'TSmiR' (Guo et al., 2014), which contained 21 miRNAs designated brain specific, we obtained a p-value of 0.145 for over-representation of 5 miRNAs in the 20 enrichment miRNAs (Fisher's exact test, see

Suppl. Table S8). Enrichment or not, many of the identified miRNAs have been found to exhibit brain enriched or specific expression (Table 3.3) or to perform regulation in brain. E.g., miR-153 is involved in Parkinson's and Alzheimer's disease (Doxakis, 2010; Liang et al., 2012), and miR-138 controls axon regeneration and has been associated to panic disorder (Liu et al., 2013; Muinos-Gimeno et al., 2011).

Some miRNAs are regulated by the transcriptional repressor REST. To uncover underlying network motifs, we assessed if there was a significant enrichment of REST-miRNAs (Section 1.2.6) in our set of enrichment miRNAs. A list of 40 REST-miRNAs (Suppl. Table S2) from Johnson and Buckley (2009), of which 22 were contained in the 153 TargetScanHuman miRNAs was compared to the enrichment miRNAs. Six out of 20 miRNAs are known or predicted to be regulated by REST (Table 3.3) and account for a significant enrichment (p-value = 0.044; Fisher's exact test). According to the ChIP-seq data of the 15 cell types, 123 miRNAs had a REST binding site within a distance of 10 kb from the TSS, among them 16 out of the 20 enrichment miRNAs. This association was, however, not significant (p-value = 1; Fisher's exact test).

Since REST is mostly expressed in non-neural tissues, one would expect that REST-miRNAs should not be expressed in those. Of possible REST-miRNAs from Table 3.3 only miR-139 had a strong expression in non-neural cell types, where REST activity is high. Thus, we assume that there is active repression of the other five known REST-miRNAs by REST.

In summary, there is evidence that a significant subset of the enrichment miRNAs can be targeted by REST whenever it is expressed, and that many enrichment miRNAs assist in the regulation of neural processes. In addition, enrichment miRNAs miR-153, miR-217 and miR-448 had predicted miRNA binding sites in the 3'UTR of REST (Figure 3.14).

Discussion

The results from this and the former chapters lead to the coherent image that REST and the enrichment miRNAs co-regulate a huge set of genes, in one or more modules, that comprises about one third of genes with neural function. Figure 3.14 gives an impression of the network set up by them. The regulation of the gene set is very complex in order to achieve cell type specific expression patterns. Below we will try to break the complexity down to simple network motifs.

The basis of the following discussion is the correctness of the identified 'detected' and 'expressed' genes. However, we have to keep in mind that it is not trivial to define when

Table 3.3: Role of enrichment miRNAs as target of REST, within neural tissue and in gioblastoma.

miRNA family	REST targets		Expressed in neural tissue[2]	Glioblastoma tumor suppressor[3]
	(Johnson and Buckley 2009)	Samples with ChIP signal[1]		
miR-101/101ab		5		
miR-129-5p/129ab-5p	known[4]	15	s,d	gs
miR-132/212/212-3p	known	15	e,s,d	
miR-138/138ab		6	s,d	gs
miR-139-5p	known	14	e,s,d	
miR-153	likely[5]	8	s,d	gs
miR-185 family		1		
miR-190/190ab		3		
miR-208 family		0	n.a.	
miR-217		1		
miR-218/218a	likely[5]	1	d	gs
miR-300/381/539-3p		1	n.a.	
miR-326/330/330-5p	miR-330 known	13	s	gs
miR-329 family		1		
miR-34 family		2	d	gs
miR-374ab		7		
miR-421		0	d	
miR-448/448-3p		0	n.a.	
miR-499-5p		0	d	
miR-543		2	n.a.	

[1]A detailed description of REST binding in proximity to the enrichment miRNAs according to ChIP-seq data can be found in Suppl. File 3.
[2](s) specific to brain (Guo et al., 2014; Sempere et al., 2004), (e) enriched in brain (Guo et al., 2014; Sempere et al., 2004), (d) detected in non-cancerous neural tissues of Landgraf et al. (2007) with a copy count of more than 3% of the total counts for all tissues.
[3]See Suppl. Table S7 for details.
[4]Known according to (Gao et al., 2012; Johnson and Buckley, 2009).
[5]Gene sequence is in the intron of REST-regulated genes. 77% of intronic miRNAs are co-expressed with their host genes (Liang et al., 2007). We assume that miRNAs situated in the introns or REST-regulated genes will be expressed with them, thus they are similarly regulated by REST.

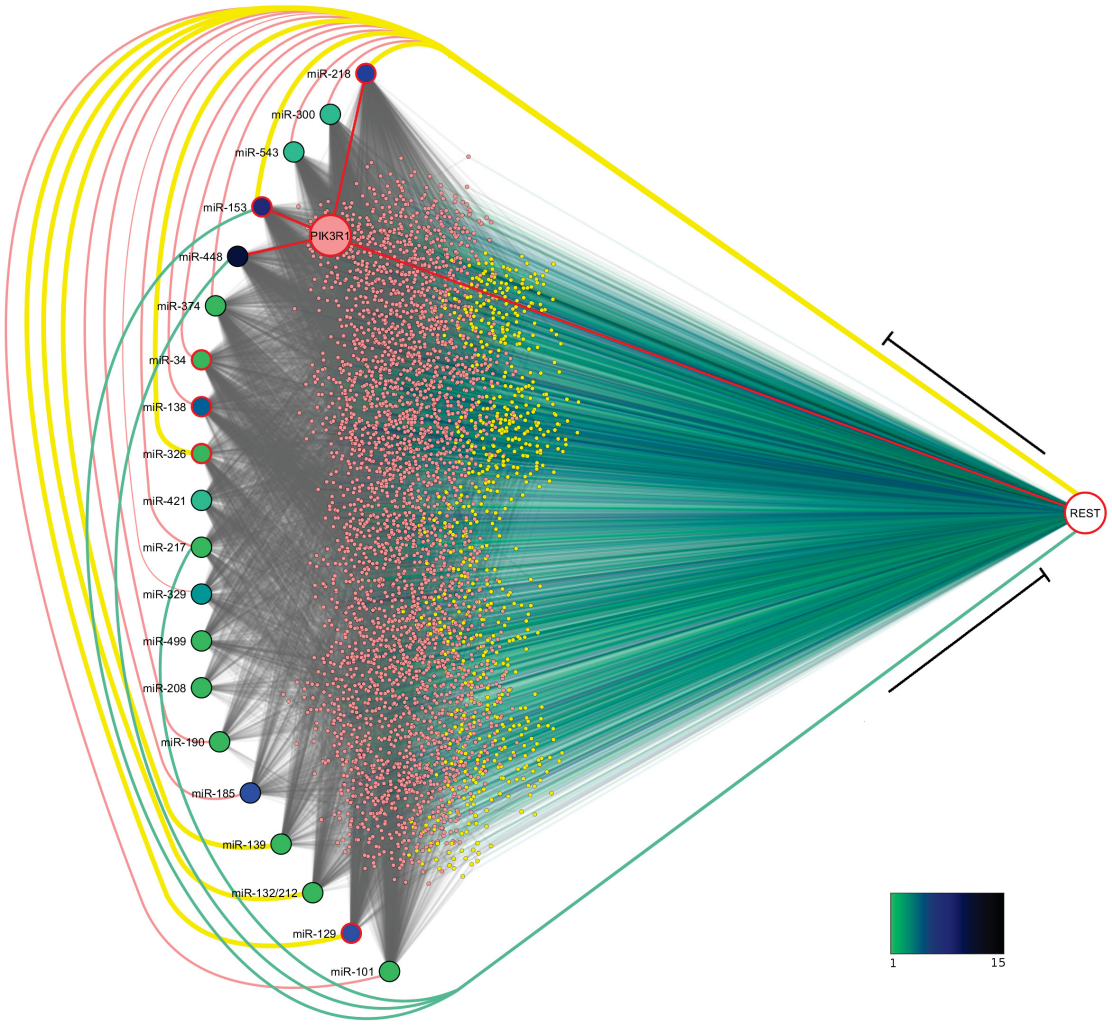

Figure 3.14: Network of regulatory interactions between REST and enrichment miRNA co-regulated genes. Yellow small circles in the center depict genes with neural function and pink circles are genes with any other function. Colored lines (hue from light green to black) show the number of tissues in which REST was detected in proximity to the respective gene. On the left there are 20 enrichment miRNAs. The circle fill color indicates the number of tissues were the miRNA was found over-represented. Predicted regulation of the genes by enrichment miRNAs is depicted in gray color. A possible regulation of enrichment miRNAs by REST is shown by means of yellow (known) or pink (deduced from ChIP-seq) curves. Green lines indicate putative regulation of REST by miR-153, miR-217 and miR-448 (TargetScanHuman 6.2). miRNAs with red borders are involved in glioblastoma and red lines connect a subset to PIK3R1 and REST. Sorting of miRNAs was done by means of hierarchical clustering performed on their number of connections to genes. Figure from (Gebhardt et al., 2014).

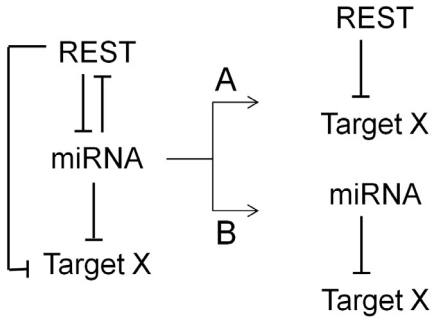

Figure 3.15: The combined I2-FFL and the double negative feed-back loops are reduced to two different 2-node motifs.

a gene or a miRNA can be regarded as 'expressed'/'detected'. E.g., very modest changes in expression level, as conveyed by miRNAs, can have fundamental effects (Bartel, 2009). We tested several cutoffs for our definition of 'expressed'/'detected', but actually every miRNA would need its own unknown expression level threshold. In addition, when using samples from different experiments and cell types, there might be batch effects in the expression level that cannot be erased by the applied normalization procedure. Hence, the above results need to be surveyed with caution.

Unfortunately in the huge repository of ENCODE (The Encode Project Consortium II) no mRNA expression data can be found for H1-neurons. As a result, it is not possible to contrast the expression level of the transcriptional repressor REST with the miRNA expression in concordant cell types. Thus, we need to refer to prior knowledge from the literature about REST expression in tissues.

REST and miRNAs can build different kinds of network motifs (Section 1.3.2). The I2-FFL can be combined with the double negative feed-back loop of REST and the miRNAs (Figure 3.15). Whenever one of the partners of the double negative feed-back loop is not there, the I2-FFL is reduced to a simple regulatory relation between either REST or the miRNAs and the respective target genes.

Figure 3.15 Condition A: REST is present
Since REST is a transcriptional repressor with the capability to completely silence its target genes, we assumed that, when it is present, most REST-miRNAs will not be expressed. We found that this is considerably true. Since REST performs repression of the common target genes, miRNA expression is not necessary and is repressed as well.

Figure 3.15 Condition B: REST is absent
REST-miRNAs operate in absence of the repressor or when its expression level is low. This is the case in neural tissue. In addition, in many neural cell types the truncated isoform REST4 is the predominant variant (see Section 1.2.1 and 1.2.5). REST4 has a reduced repressive activity (Section 1.2.1), therefore expression of the REST-miRNAs is possible here. We assume that the miRNAs perform a modest repression of the module genes to achieve precise expression levels and specific neural phenotypes. Co-targeting by multiple miRNAs is very likely here.

There are three enrichment miRNAs with predicted binding sites in the 3'UTR of REST. All of these miRNAs show quite high expression and exclusively in the H1-neurons in Figure 3.13. This observation makes it likely that these miRNAs indeed repress REST in neurons. To be sure about this relation an experimental examination of the theory would be necessary.

The function of the I2-FFL as network motif is not well understood (Section 1.3.2). From the observations above, it becomes obvious that the possibly most important function of the loop reveals only in combination with the feed-back loops. It is there to guarantee the activity of the proper set of regulators to obtain cell type specific gene expression and it is of importance beyond the definition of a single cell type.

Figure 3.15 Condition C: REST and the miRNAs are present
As we observed, there are cases in which REST and REST-miRNAs are both expressed in a tissue, e.g. miR-139 in non-neuronal cell types. Considering that REST has a low instead of a zero expression level in H1-neurons, in this cell type both repressive parts of the I2-FFL coexist. Here the I2-FFL comes in its original form.

It can only be speculated on what the miRNAs are good for, when REST is present to repress the common target genes. REST target specificity can be altered by co-factors (Section 1.2.4), therefore, an absolute repression of all module genes is not likely for all cell types. The miRNAs could assist in the repression or perform fine-tuning of the module's gene expression level. It is possible that the miRNAs need to be expressed to repress targets that are not shared with REST in the considered condition (but in another condition the gene is a REST target). In this case, the I2-FFL would be an artifact. The regulation would happen as in Condition A.

For enrichment miRNAs that are not part of the described network motifs, similar conclusions can be drawn. Enrichment miRNAs that are expressed in the cell type where they have their over-represented targets, are likely to perform important regulatory

functions. This is especially true for enrichment miRNAs in H1-neurons, where REST has a much lower repressive activity than in non-neural cell types. Here the situation is similar to Condition B when REST is regarded as not expressed or to Condition C, when it exists in low expression levels. We expect that REST suppresses common target genes when it is present, otherwise the miRNAs perform fine-tuning on their expression.

The relevance of enrichment miRNAs that are not expressed in the respective cell type where they were detected as enrichment miRNA, is much more difficult to judge. It is possible that such miRNAs are true regulators in another condition. Nice examples are miR-153 and miR-448, which can be detected as enrichment miRNAs in many non-neural cell types, but which are expressed in neural tissues. In their case, arguments for their importance in the regulatory network of H1-neurons, accumulate.

3.6.2 Enrichment miRNAs in glioblastoma - miR-448 and PIK3R1

Results

Glioma is the most prevalent malignant tumor in the brains of adult humans with glioblastomas accounting for 76% of all gliomas (Central Brain Tumor Registry of the United States , CBTRUS). REST has been shown to act as oncogene in glioblastoma in human (Conti et al., 2012; Kamal et al., 2012), we, therefore, assume that some of the REST-regulated genes must be involved in processes relevant for the disease. Since the identified enrichment miRNAs have functions in neural tissues and numerous connections to the transcriptional repressor, we expect to find some of them being involved in glioblastoma as well. Many studies have been devoted to the analysis of miRNA impact on glioblastoma and identified several miRNAs with properties as tumor suppressor (see Suppl. Table S7). Interestingly we found glioblastoma tumor suppressor miRNAs enriched (p-value = 0.018, Fisher's exact test) in the high confidence subset of six enrichment miRNAs that we defined as miRNAs with significant over-representation in more than 5 cell types (Table 3.2 and Table 3.3).

The four glioblastoma related miRNAs were miR-129, miR-138, miR-153 and miR-218. The remaining two miRNAs from the high confidence subset miR-185 and miR-448 have not been identified as glioblastoma tumor suppressors. Due to their proximity to the known tumor suppressors in the network of genes regulated by REST and by the enrichment miRNAs we found it likely that they would function as tumor suppressors, too. Particularly, miR-448 was of interest because it was over-represented in 13 of the 15 analyzed cell types (Table 3.2), which is the highest number in the total analysis. We

found that it was expressed in H1-neurons (Figure 3.13). There is nothing known about the expression of miR-448 in glial cells or about its function in neural tissue in general.

miR-448 and PIK3R1:
We next wanted to illustrate the potential value of our predictions for discovery of biologically relevant knowledge. To further examine a possible function of regulatory feed-back loops between REST and enrichment miRNAs in glioblastoma, we looked for candidate genes implicated in this disease that were REST targets with predicted binding sites for miR-185 or miR-448.

An interesting gene fulfilling these conditions was the oncogene PIK3R1, which is known to promote proliferation and invasiveness in glioblastoma (Weber et al., 2011). The PIK3R1 transcript has two predicted binding sites for miR-448 and appeared suitable for our purpose because it also exhibits binding sites for the two known tumor suppressors miR-153 and miR-218. We tested the effect of miR-448 on PIK3R1 expression *in vitro* by means of a miRNA-UTR-assay, using a reporter plasmid with two long fragments cloned into the 3'UTR, one for each miR-448 binding site (Figure 3.2, see Section 3.2.5 *miRNA-UTR-assay*).

Figure 3.16 shows that there was a significant reduction of relative luciferase activity of the reporter plasmid with Region B of the PIK3R1 3'UTR included in comparison to control samples. This reduction could not be seen for Region A or for Region B when the miR-448 binding site was deleted. Hence, we showed that one of the miR-448 binding sites in the 3'UTR of PIK3R1 can be targeted by miR-448 with a repressive effect on gene expression *in vitro*.

Discussion

Despite the found *in vitro* impact of miR-448 we cannot make statements about whether the regulatory relation between miR-448 and PIK3R1 exists *in vivo* and particularly if this is relevant for glioblastoma. The miRNA typical way of action to co-operate in repression (Section 1.3) argues for a true relation. It is conceivable that miR-448 is one of the many miRNAs that act in concert to control cell cycle and proliferation in glial cells and other neural tissue. The high number of miRNAs identified as tumor suppressors in glioblastoma illustrates the importance of the miRNA regulatory level for smooth cell cycles and proper development of specific neural cell types. Many miRNAs make a contribution to these processes and the search for enrichment miRNAs is a suitable tool to identify new candidates and make suggestions for miRNA functions that can be tested experimentally. Even if miR-448 was not involved in glioblastoma, a true function in

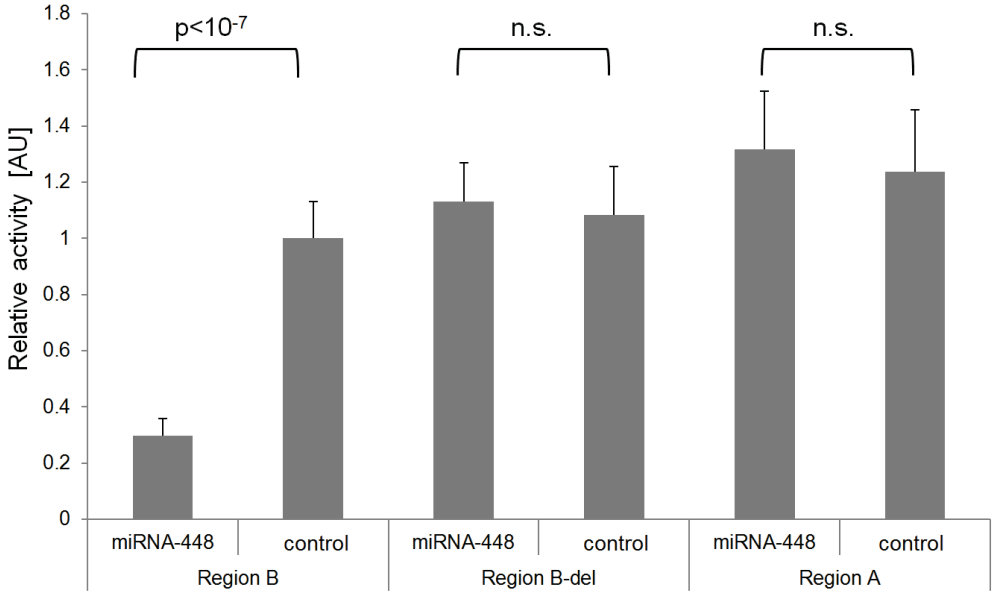

Figure 3.16: The impact of miR-448 transfection on HEK293 cells with a dual reporter plasmid containing two ∼850 bps long fragments with each of the miR-448 binding sites of the PIK3R1 3'UTR. Measurements are reported as mean of relative luciferase activity (renilla/firefly). All data are scaled so that the mean of control for Region B was 1.0. The standard deviation is shown in error bars. AU means absorbance units.

neural tissues is very likely due to its expression in H1-neurons, due to the similarity in target gene sets and expression to other miRNAs with known neural function, and due to the strong interrelation with the transcriptional repressor REST. Thus, the search for enrichment miRNAs can shed light on miRNA function.

As a caveat, we must note that when we infer the function of miR-448 by similarity of targets with other miRNAs, we are dependent on correctly identified miRNA functions of the other tumor suppressor miRNAs and on properly identified regulatory relations between miR-448 and the underlying REST-regulated network (Figure 3.14). One has to keep in mind that the 29 glioblastoma tumor suppressor miRNAs stem from different non-neoplastic references and experimental setups that cannot be compared easily (Visani et al., 2013). Moreover, all limitations and possible errors mentioned in Section3.4 and 3.5 also have impact on this kind of approach.

3.7 Extension of the approach

3.7.1 Integrating expression data and motif search

Results

As mentioned earlier, accurate identification of target genes from ChIP-seq data is challenging. We wanted to find out if we can improve the results of our methods by applying gene lists with higher fractions of true positives.

One common way to increase the fraction of true positives is the integration of expression data. Arnold et al. (2012) published ChIP-seq data for REST in combination with mRNA expression profiles before and after knock-out of the transcriptional repressor in mouse (see Section 2.2.1). It is possible to generate a new gene list from these data comprising genes with ChIP-seq peaks and differential expression patterns in NPs (94 genes).

Another way to reduce the number of false positive peaks, before peak-gene association is performed, is to search for a transcription factor binding motif (RE1) inside the peak region. This was done to generate a second gene list enriched in true positives (261 genes, see Section 3.2.6 *Integrating expression data and motif search*).

We performed a search for enrichment miRNAs on the two gene lists and contrasted them with results from the gene set produced from ChIP-seq data without this refinement (405 genes, all gene lists in Suppl. File 4). The probability distribution of the FDRs is presented as negative logarithm to base 10 in Figure 3.17. The left image shows the density distribution across the whole range of FDRs and the right figure zooms into the region of FDR 0.01 to 0.1, which equals a -$\log_{10}$(FDR) of 1.0 to 2.0.

In general, comparing two gene sets to each other, one would expect higher FDRs for the better set and a clearer separation into significant and nonsignificant results. The unrefined gene list had the highest amount of miRNAs with FDR = 1.0 (-$\log_{10}$ = 0.0) and many miRNAs that passed the significance threshold of FDR = 0.1. It clearly outperformed a gene list generated from ChIP-seq data in combination with differential expression data. Refining the gene list by means of motif search yielded a reduction of FDRs with a value of 1.0 and only a modest improvement towards significant FDRs.

Discussion

For the transcriptional repressor REST, integration of mRNA expression data does not improve the outcome of the search for enrichment miRNAs. This might be due to the

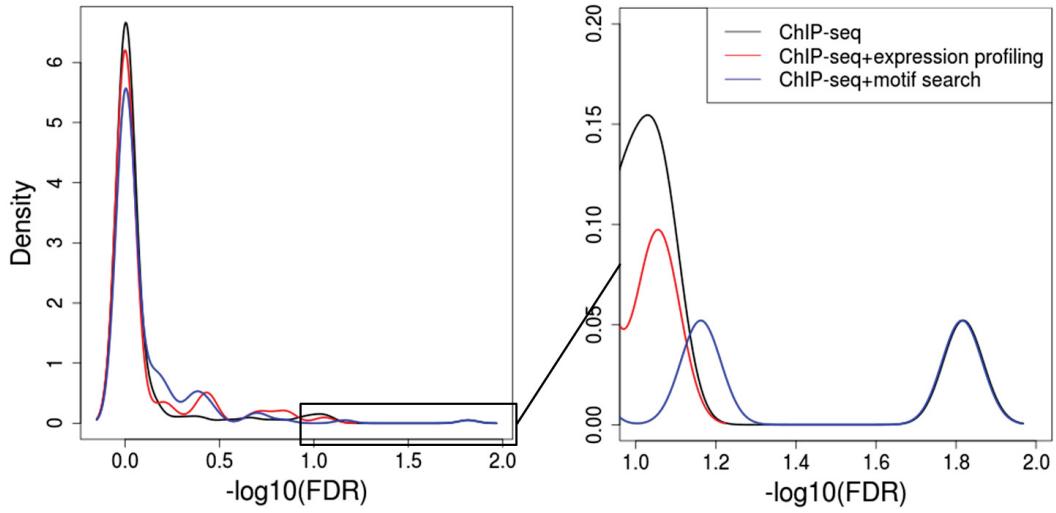

Figure 3.17: Comparison of FDRs from search for enrichment miRNAs with three different gene lists (data for NPs):
1. From ChIP-seq data obtained by peak-gene association with ranked method (black curve).
2. Intersection of list 1 with genes up-regulated after knock-out of the transcriptional repressor REST, according to mRNA expression profiles (red curve).
3. Genes from list 1 with RE1 binding motif inside the peak region (blue curve).

fact, that by restricting the applied gene set to REST-regulated genes from a single cell type the full power of the analysis is not utilized. REST occupies a huge amount of binding sites that do not become de-repressed when the factor is knocked out (Otto et al., 2007), in our case 94 out of 405 candidate genes, but many of these binding sites are authentic. On the one hand, it is possible that they are not de-repressed because more is needed than the loss of the transcriptional repressor to reverse repressive epigenetic marks (see Section 1.2.4). On the other hand, the binding sites might be inactive, e.g. due to a missing co-factor. Thus, filtering the ChIP-seq gene list with mRNA expression data only makes sense when a scientific question related to a specific cell type and condition is to be answered. E.g., by searching for enrichment miRNAs on a list of target genes differentially expressed in the presence of an activator, one could identify miRNAs that are part of a I1-FFL with the activator. REST and its enrichment miRNAs share many targets but predominantly not in the same cell type, as we learned in Section 3.6.1. In our approach, we intended to gain an insight into a system-wide view of the regulatory network of REST and miRNAs; thus, it is not reasonable to limit the REST target gene list to a certain condition.

Filtering the ChIP-seq peaks by means of the binding motif before the peak-gene association step yields an accurate list of REST targets (assuming that the peak-gene association works well). However, one has to be aware that genes containing a canonical RE1 motif tend to be targeted by REST in many cell types; more specific targets are not recorded (see Section 1.2.2 and 3.3.1). The fraction of these 'common' REST targets is higher in the filtered than in the unfiltered list. This means that actually a subnetwork of 'common' REST targets and miRNAs is analyzed.

In summary, with each gene list a different scientific question can be tackled. Additional filtering of ChIP-seq generated lists of REST-regulated genes does, however, not seem to be helpful for our experiments.

3.7.2 Integrating DHS sites

Results

ChIP-seq is not the only method that yields genome-wide information on transcription factor binding. In the following section we want to compare its performance to an alternative method. Regions hypersensitive to digestion by DNase I or open Chromatin regions show low nucleosome density and are often targeted by active gene regulatory elements (Wu and Gilbert, 1981). Nowadays it is possible to construct genome-wide maps of DHS sites by means of next-generation sequencing (Song et al., 2011). For eight of the 15 cell lines with ChIP-seq on REST there were DHS site profiles available. We first examined to what extent REST ChIP signals were observed inside or outside of DHS regions (see Section 3.2.6 *Integrating DHS sites*, Suppl. Table S9) and found that the majority of REST peaks was contained in DHS regions, on average 71.5%.

To test how application of DHS data can assist in the search for enrichment miRNAs, we contrasted the results of our simulation approach on three gene lists: (i) genes found by ChIP-seq, (ii) genes found with a combination of ChIP-seq and DHS regions, and (iii) genes found by means of RE1 motif search in DHS regions. We did this for six out of the eight cell types focusing on those with the highest amount of significant results. The probability distributions of the FDRs are presented as negative logarithm to base 10 in Suppl. Figure S6. They show that there were cell types, e.g. ECC1, in which the miRNAs with the best FDRs could be obtained by using the gene list produced without ChIP-seq data (list iii). To compare the results in another way, we took the top 10 -$\log_{10}$FDRs from two samples each and computed p-values with a one-sided Wilcoxon rank-sum test. In Figure 3.18 we show the top 10 -$\log_{10}$FDRs of the three approaches next to each other. The p-values for the partners with higher -$\log_{10}$FDR

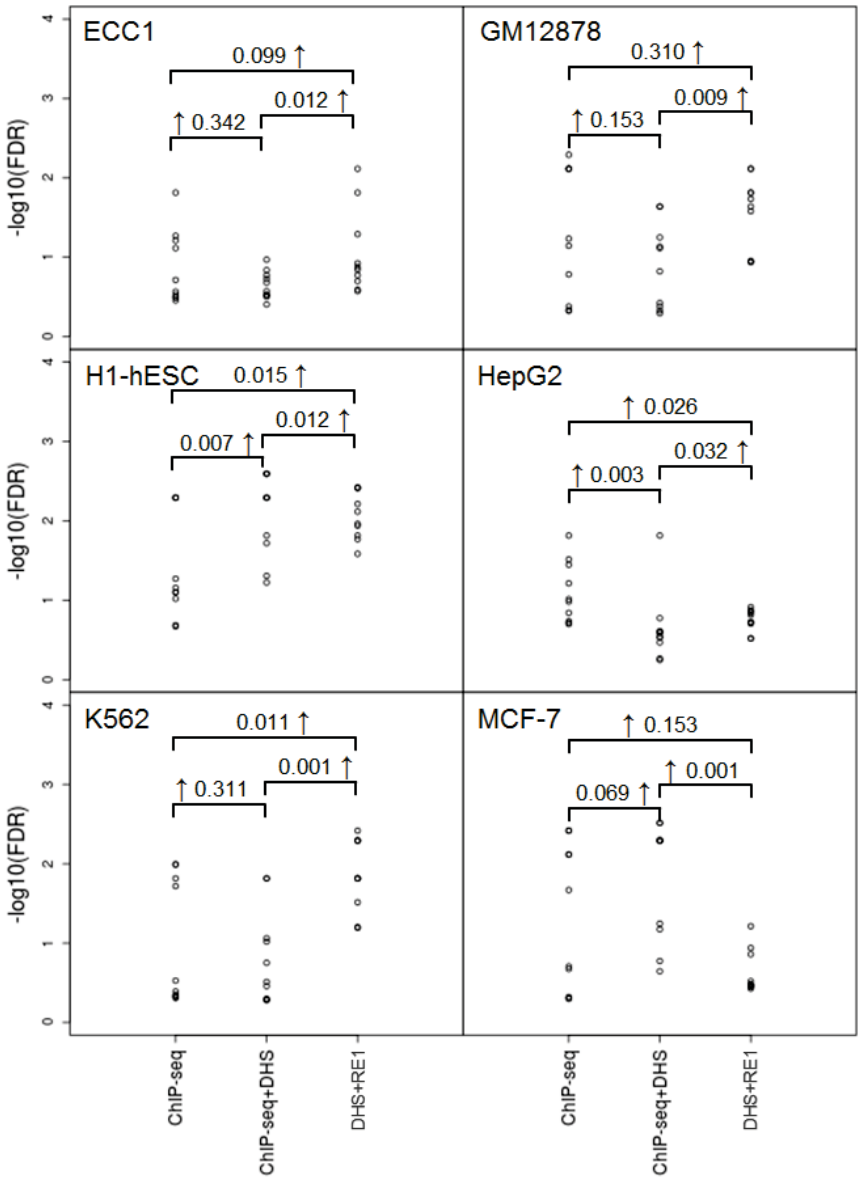

Figure 3.18: The ten highest -$\log_{10}$FDRs of miRNAs were plotted for three gene lists per cell type:
1. From ChIP-seq data.
2. From regions of overlapping ChIP and DHS signal.
3. From regions of open chromatin with RE1 motif.

Results of a one-sided Wilcoxon rank-sum test are presented above the values. The position of the arrow left or right of the p-value designates the set with higher -$\log_{10}$FDR values.

values are marked with an arrow on the respective side on the p-value. The top 10 sets are presented for each of six cell types.

In four out of six cell types (ECC1, GM12878, H1-hESC and K562) using only DHS regions with a motif filter yielded higher values for -$\log_{10}$FDR, meaning lower FDRs than any ChIP-seq including method. New enrichment miRNAs were suggested, miR-125, miR-128, miR-149, miR-214 and miR-24, and miRNAs found before, were confirmed. All miRNAs with a FDR < 0.1 are listed in Suppl. Table S10. Again in four out of six cases (ECC1, GM12878, HepG2 and K562) ChIP-seq data alone performed better than an overlap of ChIP and DHS signal.

Discussion

The question of which method to prefer for generation of gene lists for the search for enrichment miRNAs cannot be answered with certainty because all three methods can yield significant results and their differences are small (Figure 3.18). For factors with short sequence motifs we expect that DHS regions with motif filter cannot perform as well as in the case of the transcriptional repressor REST due to the introduction of more false positive binding sites. Nevertheless, whenever ChIP-seq and DHS data are available it is probably advisable to make use of them separately and to compare the result in the end. It is conceivable that each data type helps to uncover another part of the underlying regulatory network.

3.7.3 Application on other factors

Results

The search for enrichment miRNAs can be performed for other DNA binding proteins than REST. Here we focus on lists of genes regulated by either an activator or a repressor. To demonstrate that the procedure works for other factors, we used ChIP-seq data originating from experiments with the HepG2 cell line from the ENCODE project for many different factors (The Encode Project Consortium I, Suppl. Table S1). We performed the experiments as described above with the REST ChIP-seq datasets. The results are presented in Table 3.4.

Table 3.4: Identified enrichment miRNAs from ENCODE ChIP-seq data in HepG2 and with various antibodies.

Target/ Antibody	Description	Nr. targets/ Nr. in TargetScanHuman	Enrichment miRNA (p-value)
ATF3	Activating Transcription Factor 3, activator and repressor of transcription	779/514	-
CEBPB	C/EBP_Beat, activator of transcription	3364/2285	-
CREB1 (SC-240)	CAMP-Responsive Element-Binding, cAMP dependent transcriptional activator	4026/2672	-
CTCF (SC-5916)	CCCTC-Binding Factor, transcriptional regulator binds (HAT)- or (HDAC)-containing complexes, activator and repressor	7956/4871	-
ELF1 (SC-631)	E74-like Factor 1, activator of transcription	8545/5497	miR-142-3p (0.008) miR-7/7ab (0.008)
FOXA1 (SC-101058)	Forkhead Box A1, transcriptional activator	6825/4530	miR-205/205ab (0.015)
HDAC2 (SC-6296)	Histone Deacetylase 2, forms transcriptional repressing complexes	2250/1597	-
Max	MYC Associated Factor X, can be transcriptional activator (MYC-MAX) and repressor (MAD-MAX)	6312/4259	miR-205/205ab (0.038) miR-93 family (0.031)
MYBL2 (SC-81192)	V-Myb Myeloblastosis Viral Ongcogene Homolog, activator and repressor of transcription	8926/5673	miR-140 family (0.036) miR-142-3p (0.015) miR-1ab/206/613 (0.008)
NR2F2 (SC-271940)	Nuclear Receptor Subfamily 2, ligand inducible transcription factor	801/552	-
p300	E1A Binding Protein P300, co-activator, HAT (histone acetyltransferase)	3757/2552	miR-142-3p (0.061) miR-144 (0.084) miR-194 (0.099)

- *Continued on next page* -

Table 3.4 – *Continued from previous page*

Target/ Antibody	Description	Nr. targets/ Nr. in TargetScanHuman	Enrichment miRNA (p-value)
Pol2	DNA-Directed RNA Polymerase II	10791/6839	miR-124/124ab/506 (0.034) miR-140 family (0.061) miR-142-3p (0.015) miR-1ab/206/613 (0.008) miR-200bc/429/548a (0.015) miR-205/205ab (0.073) miR-7/7ab (0.015) miR-93 family (0.036)
Sin3Ak-20	SIN3 Transcription Regulator Homolog A, transcriptional repressor, co-repressor of REST	4001/2317	miR-1ab/206/613 (0.046)
SRF	Serum Response Factor, transcription factor	1191/789	-
USF1 (SC-8983)	Upstream Transcription Factor 1, transcription factor, activator	4640/2801	-
YY1 (SC-281)	YY1 Transcription factor, activator and repressor of transcription	5380/3574	miR-300/381/539-3p (0.092)
ZEB1 (SC-25388)	Zinc Finger E-Box Binding Homeobox 1, repressor of transcription	404/268	miR-153 (0.015) miR-448/448-3p (0.015)

Enrichment miRNAs can be found for many, but not all factors, both for repressors and activators. In contrast to our findings for cell line A549, we do not find dependencies on the size of the gene set because results can be obtained either for very large sets such as from MAX or small datasets such as from ZEB1. We found miR-1 with enriched targets in SIN3A regulated genes; notably, although SIN3A is a co-repressor for REST (Huang et al., 1999), miR-1 was not among the enrichment miRNAs of REST.

Discussion

It was to be expected that there would be factors without enrichment miRNAs. These do not co-operate with miRNAs to an extent that could be captured by our method, at least under the given conditions. For repressors, the same considerations with respect

to network motifs and manner of co-operation with the miRNAs apply, as for REST. Regarding activators for which we found enrichment miRNAs another network motif must be considered. Activators, miRNAs and target genes build an I1-FFL (Figure 3.19).

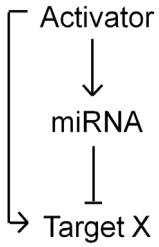

Figure 3.19: An activator, a miRNA and their common target genes span an I1-FFL.

According to Alon (2007), the I1-FFL can generate a pulse and can act as response accelerator. On the one hand, the activator stimulates the expression of the common target X. On the other hand, by triggering the miRNA expression, the target is soon repressed to a repression threshold, resulting in a peak in target X production. The network motif allows to reach a steady state very quickly (Alon, 2007). In contrast to the I2-FFL, the essential function of the I1-FFL can be observed within single cells. That is why it was possible to determine the function of such motif by means of single cell organisms such as *Escherichia coli* (e.g. in Ronen et al., 2002).

Some of the tested factors for which we observed enrichment miRNAs had large gene sets. As we learned earlier (Section 3.4), the over-representation of enrichment miRNA targets in the gene list is more difficult to detect for large datasets, because they are too similar to the random background. In such cases, observing a significant over-representation must be considered even more relevant.

In Table 3.4 we also mention enrichment miRNAs found using ChIP-seq data of DNA Polymerase 2, which is neither activator nor repressor. Theoretically the analysis can be performed with any gene list and can bring useful findings. However, the results will not be obtained due to similarities or interrelations in regulatory effects of a certain transcription factor and co-operating miRNAs and must be due to some other correlation whose biological meaning currently escapes us.

3.8 Does filtering work?

After demonstrating that REST and miRNAs span an extensive regulatory network around common modules and observing the existence of these modules for other factors, we now want to answer the central question of the thesis. Can ChIP-seq data be used to improve miRNA-target predictions? Are the targets of enrichment miRNAs better predicted than other targets?

Results

The 20 enrichment miRNAs together with REST are possible regulators of a set of 3,814 genes (34,2% out of 11,161) in form of 8,438 miRNA-gene associations, which we refer to as 'filtered' set, (11.6% out of 72,770 miRNA-gene associations, Figure 3.14, Suppl. File 5). We hypothesized that the enrichment of predicted miRNA targets in the REST target gene lists points to a higher number of true miRNA-target predictions in the filtered set in respect to the total background. This enrichment in true relations would originate from the existence of groups of genes that need to be repressed both on pre- and post-transcriptional level in a coordinated way.

To test the hypothesis, we checked if a significant enrichment of experimentally proven miRNA-target associations in the filtered set in respect to the background could be obtained (see Section 3.2.4 *Significance of filtering miRNA target predictions*). We applied data from TarBase 6.0 as source of validated miRNA-target pairs. The result of the analysis can be found in Table 3.5.

Five miRNAs had more than 10 predicted miRNA-target associations and were con-

Table 3.5: Significance of enrichment of valid miRNA-target associations in the filtered set. Table adapted from (Gebhardt et al., 2014).

miRNA family	All/ validated pairs	Filtered/ filtered validated pairs	Proportion valid all [%]	Proportion valid filtered [%]	Fold enrichment	*p*-value
miR-101/101ab	804/65	635/50	8.08	7.87	0.97	0.726
miR-132/212/212-3p	407/25	332/21	6.14	6.33	1.03	0.498
miR-218/218a	931/16	746/16	1.72	2.14	1.25	0.028
miR-34 family	680/43	500/36	6.32	7.20	1.14	0.078
miR-374ab	656/11	530/11	1.68	2.08	1.24	0.094
merged data	3478/160	2743/134	4.60	4.89	1.06	0.071

sidered in the analysis. For four of them there was a fold enrichment larger than 1, with miR-218 yielding the best fold enrichment of 1.25 and a p-value of 0.028. Notably, this was a miRNA that was over-represented in 10 out of 14 cell types. A mildly significant fold enrichment was obtained for the union of all five considered miRNAs.

Note that the modest fold enrichments are to be expected since we are comparing a subset of miRNA targets with a background that might contain many true positives. The direction of the fold change is indication enough of the significance of our analysis.

To challenge our hypothesis in another way, we consulted a feature of miRNAs that we had not used before. Grimson et al. (2007) found that miRNA binding sites that are in close proximity (8 to 40 bps), so called 'dual sites', often co-operate in down-regulation of a target gene. We assumed that, if the filtered gene set had a higher proportion of valid targets, it would be rich in 'dual sites' in comparison to a random background. And this is what we found with the help of 1,000 test cases. Each test yielded a sum S of 'dual sites' for 123 randomly selected genes (see Section 3.2.4 *Significance of filtering miRNA target predictions*). When all S values of the filtered set were compared to all S values of the background by a one sided Wilcoxon rank-sum test, the portion of the filtered set turned out to be greater with high significance (p-value $= 1.14 \cdot 10^{-5}$). Strikingly, this result was obtained although we restricted the analysis to miRNA binding site classes with equal or lower miRNA binding site density in the filtered set compared to the random background. E.g., for class $i = 2$ the 3'UTR length distribution was significantly shifted towards longer 3'UTRs as compared to the background (p-value $=$ 0.072) leading to a reduced likelihood in the filtered set to find 'dual sites' by chance. The difference is modest but visible in Figure 3.20.

Discussion

The accumulation of true positive miRNA targets as well as of 'dual sites' in the REST target genes argue for the hypothesis that miRNA binding site predictions can be filtered due to the presence of common modules controlled by two regulatory levels, the pre- and the post-transcriptional, and that ChIP-seq data can be used to gain knowledge about regulatory relations on the miRNA level. It has to be stated that the filtering is not a clear sorting for 'yes' and 'no'. Instead, we obtain a subset of predicted miRNA targets that have an increased likelihood of being regulated by the respective enrichment miRNAs. This can be useful when a scientist needs to select candidate interactions for experiments, as we illustrated with our *in vitro* testing of the effect of miR-448 on PIK3R1.

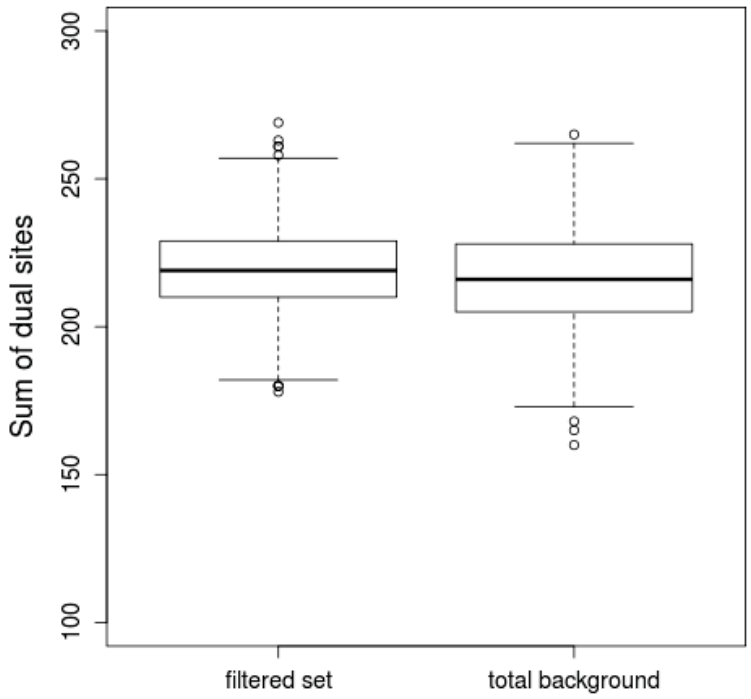

Figure 3.20: After selection of 123 random genes from filtered set and background set, the number of 'dual sites' was determined. The procedure was repeated 1,000 times and the numbers of 'dual sites' were summarized in the boxplots. The number of 'dual sites' was higher in the filtered miRNA-target predictions.

The relation will need a deeper validation in the future, as soon as more experimentally validated miRNA-target associations are available. It would be nice to further confirm the results with a solid data basis and for other factors than REST.

The condition-independent character of the identified relations, while being useful for the establishment of the general regulatory network, brings a major limitation of the approach. We can make statements about miRNAs that have impact on certain genes with a high likelihood, but we cannot give information about where the interaction takes place. This question can hardly be solved with the help of experiments, considering the huge number of different cell types and conditions in complex organisms such as human and mouse. We expect that such information will eventually be found computationally by means of network analyses.

3.9 mBISON web application

A crucial advantage of our simulation approach over former methods to assess over-representation of miRNA targets in gene lists, is the short computing time. It allowed to set up an application that can be accessed online and calculates significance of enrichment in a reasonable time period (see Section 3.2.7). The mBISON (miRNA binding site over-representation) application was designed to be used by all researchers even without experience in bioinformatics. It can be used for human and mouse data and custom background sets can be supplied.

To be able to make use of the ever growing amount of ChIP-seq data deposited in databases such as GEO (Edgar et al., 2002), we made the direct use of such data in BED-format possible. Up to three files can be uploaded. On the web page it is possible to perform peak-gene association with a choice of methods (binary method with 5, 10 or 20 kb, ranked and strict ranked method, see Chapter 2). Due to extensive computational demands, it was unfortunately not possible to implement the 'Ouyang' and 'ClosestGene' methods in the web application. Instead, there is a hint on the web page that recommends and refers to the R package from Sikora-Wohlfeld et al. (2013). If two or three BED-files are uploaded, only genes identified at least twice by the peak-gene association method will be considered. Subsequently, the gene list can be forwarded directly to mBISON.

Figure 3.21 shows the web appearance of the mBISON application on `http://cbdm.mdc-berlin.de/~mgebhardt/cgi-bin/mbison/home/`.

mBISON will run either with a freshly generated gene list from the peak-gene association tool or with an uploaded gene list. The user has the choice between different identifiers (Entrez ID, Gene Symbol, EnsemblGene ID or RefSeq ID) but Entrez IDs are recommended.

We restricted the size of the gene list to minimally 20 and maximally 6000 genes because, on the one hand, we found that too large gene lists usually do not yield significant results (only 11,161 genes in the background for human and 9,075 for mouse) and, on the other hand, we want to avoid extensive run times of the queries. The user can choose the number of randomizations for p-value calculation (1,000, 10,000 or 100,000).

We also allow defining a cutoff in the minimal number of target genes, that an enrichment miRNA must have in the gene list analyzed, in order to be considered for analysis. This functionality was implemented to avoid that miRNAs with very small numbers of predicted binding sites appear in the results by chance. If very small gene lists are supplied, this number should be set to zero.

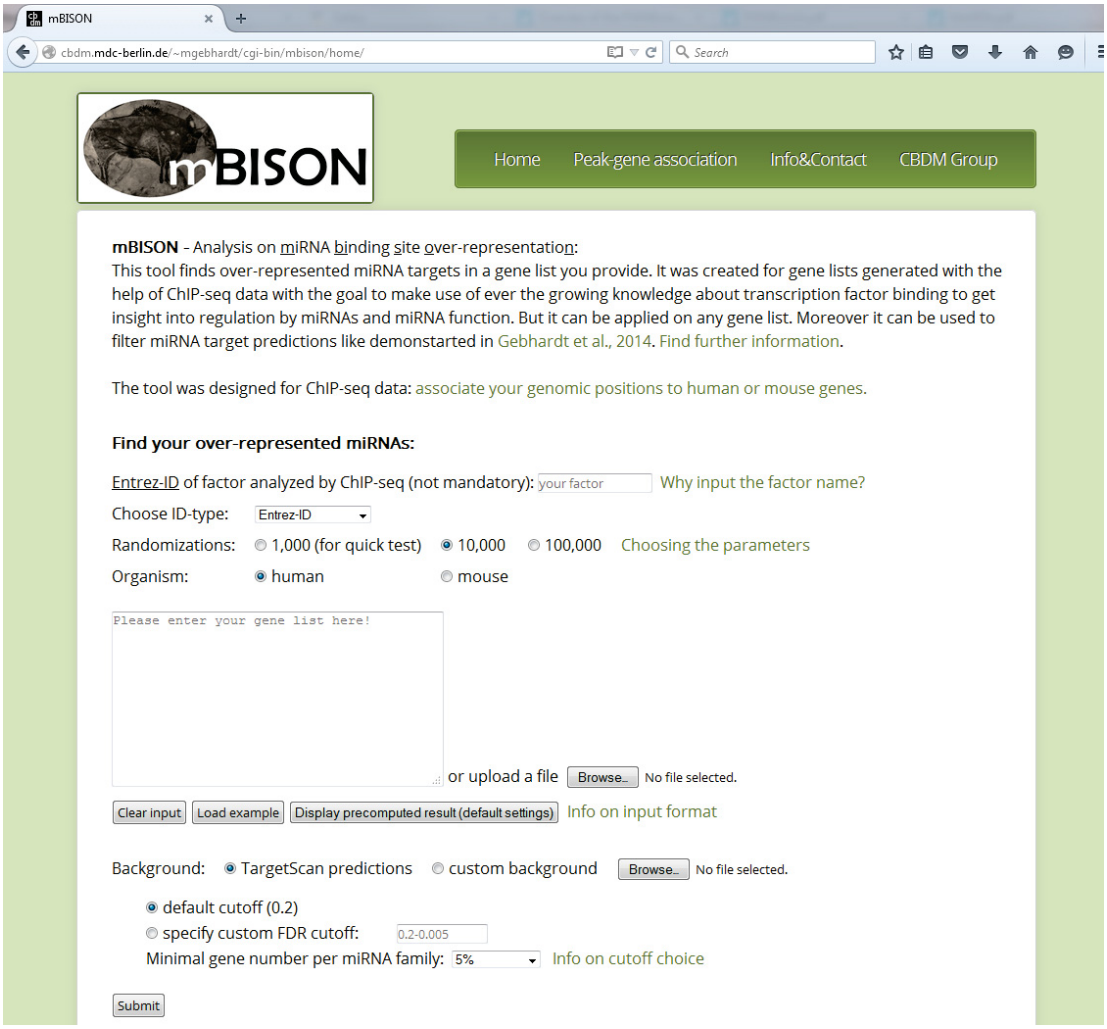

Figure 3.21: Home page of the mBISON web application.

For the reporting of results of enrichment miRNAs, a significance threshold for the FDR can be chosen (from 0.2 to 0.005).

The mBISON output comprises a table of p-values and FDRs for all 153 miRNAs, which is sorted by FDR. Optionally, if the user provides the Entrez ID of the master factor that was used for the ChIP-seq experiment (e.g., 5978 for REST), miRNAs with predicted binding sites in the 3'UTR of the master factor will be listed. If ChIP-seq data were the input (but not in case of gene lists) miRNAs in proximity to ChIP signals will be mentioned, meaning miRNAs with a peak in distance of up to 10 kb (according to miRBase, release 20; Kozomara and Griffiths-Jones, 2011). All this information can

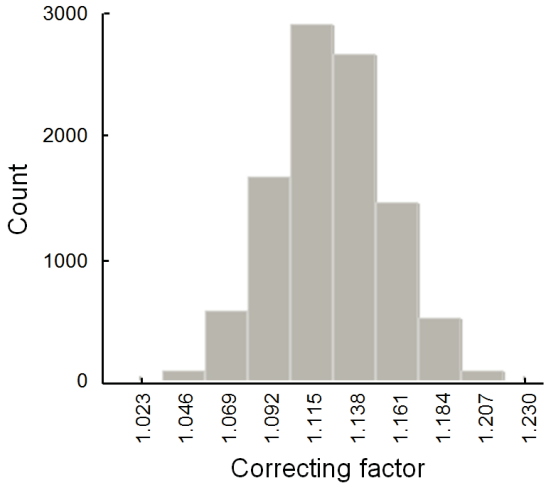

Figure 3.22: A part of the mBISON output regards the correcting factor statistics, which give information on the average number of predicted miRNA-gene pairs of the query gene set.

be used to identify feed-back and feed-forward loops.

Another part of the output regards the correcting factor statistics. A histogram of all correcting factors used in the analysis can be viewed and interpreted (Figure 3.22). The average correcting factor will be larger than one, if there are more miRNA-gene interactions predicted for the 3'UTRs of the input gene set than for the average background.

The described output can be viewed on the web page or can be downloaded in TXT-format. In addition, all miRNA-gene pairs from gene set and enrichment miRNAs can be obtained for subsequent analysis.

The mBISON application was published in 2015 (Gebhardt et al., 2015). The application is designed to identify enrichment miRNAs and network motifs, but it can be used to improve miRNA-target predictions as demonstrated in Section 3.8.

3.10 Conclusion

Since it is very difficult to detect miRNA binding sites in the 3'UTR of genes experimentally on a genome-wide scale, scientists are often dependent on miRNA-target predictions that suffer from high rates of false positives. In contrast, nowadays there is a huge amount of experimental data on transcription factor binding. miRNAs and transcription factors can co-operate in gene regulation. If the regulation is not restricted to a single gene, but covers sets of genes or modules, it becomes conceivable that knowledge on miRNA-targets can be inferred from transcription factor binding information. We set out to examine if ChIP-seq data on a specific transcription factor, in our case REST, can be used to detect miRNA-target predictions of biological relevance.

We found evidence that such detection is possible if the pre- and post-transcriptional level of regulation of a gene set are sufficiently interwoven. With our approach, we were able to study the targets of the transcriptional repressor REST, their cell type specificity, their potential post-transcriptional regulation and their biological function, which accordingly defines a function for the regulating miRNAs. This was achieved without the need of extensive miRNA profiling by means of a search of miRNAs with enriched targets in lists of REST-regulated genes, which were obtained from ChIP-seq data, and on the basis of a comprehensive collection of knowledge about the transcriptional repressor. Due to the application of many different cell types it was possible to view the global regulatory network of REST and the enrichment miRNAs in a condition-independent manner, and although it is doubtlessly not complete, we obtained an impression of how the network enables cell type specific regulation of target genes. Obviously a combination of a I2-FFL and double-negative feed-back loops assists with the formation and maintenance of particular cell types, assigning the I2-FFL a previously unknown function.

The applied randomization approach made our method reliable and fast, so that it became feasible to incorporate it into a web application. This web tool makes our method accessible to everyone, bioinformaticist or not. The application can be helpful in every field of biology and medicine since miRNAs are involved in almost all biological processes. As illustrated with our *in vitro* analysis of the effect of miR-448 on PIK3R1, the results can assist in the generation of new hypothesis and in focusing research on relations with high biological significance.

We expect our approach to improve in the future for multiple reasons, mainly with the

expected increasing number of validated miRNA targets. Improvements in the prediction of miRNA targets will also improve our method.

A clearer interpretation of the results of our method will be possible as soon as ChIP-seq data are available that are produced with an antibody directed against single isoforms of a transcription factor.

Moreover, it was shown in the thesis that ChIP-seq data do not need to be the only source of information on transcription factor binding activity. We found it highly useful to apply DHS data for the enrichment analysis, if available. We expect that the addition of other types of cell-specific data to complement the ChIP-seq data will improve the association of peaks and thus the quality of the results. We observed, that expression information was not very helpful, since the correct detection of the physical binding of the factor seems to be more relevant for the evaluation of condition unspecific regulatory effects, than whether it has an effect on expression or not. Thus, we imagine that certain types of data regarding the 3D-structure of the genome, such as the position of topologically associating domains from Hi-C contact data (Belton, 2012) might be integrated in the analysis.

The expression of a protein is influenced by so many factors, epigenetically, on the pre- and post-transcriptional level, again on level of translation and even afterwards through modes of degradation and protection, that it can hardly be grasped by the human brain. We need to break the regulation down to simpler processes, that we can understand as network motifs and in the most possible detail. Our work is only a step in this direction, but we believe that from the view of our analyses particular rules governing the global gene regulatory network and its impact on phenotypes becomes seizable. Slowly but surely, molecular biology is going to the next level.

3.11 Contributions

I did all the computational analyses and interpretation of the results in this chapter under the supervision and with the support of Prof. Miguel Andrade. The miRNA-UTR-assay with miR-448 targeting of PIK3R1 3'UTR was performed by Stefanie Reuter and Ralf Mrowka.

Supplementary Information

S1 Supplementary Methods

General Methods

This section describes methods and gives information, which were used in more than one experiment or are very general.

Analysis on differentially expressed transcripts

Microarray data were processed using R statistical programming language (http://www.r-project.org, R Development Core Team) and the libraries *limma* (Ritchie et al., 2015) and *affy* (Gautier et al., 2004). Normalization was performed by means of the Robust Multi-array Average (RMA) approach applying the rma()-function. Probe annotation was provided by the *mouse4302.db* and *hgu133plus2.db* for mouse and human, respectively. The Benjamini and Hochberg method helped to correct FDRs for multiple testing (Benjamini and Hochbert, 1995). Cutoffs were chosen depending on the experimental setup and are specified in the corresponding subsections.

Source of genomic positions of RefSeq transcripts

Genome assembly mm9 and hg19 were used for mouse and human, respectively. The genomic positions of all known RefSeq transcripts on natural chromosomes were extracted (Suppl. Table S1). Afterwards allocation of the RefSeq transcripts to the appropriate Entrez genes was possible. If more than one transcript was available for a gene, the longest transcript was picked. The resulting list of genomic locations for Entrez Genes with each one Entrez Gene ID was the basis for all peak-gene association procedures conducted in the subsequent experiments.

Finding overlaps of genomic positions

R statistical programming language (R Development Core Team) with the functions GRanges() and findOverlaps(subject=x,query=y,type="any",select="all") from the *GenomicFeatures* library (Lawrence et al., 2013) and its dependencies were applied for the detection of overlaps of genomic positions from ChIP-seq experiments with RefSeq genes, from ChIP-seq peaks with other ChIP-seq peaks and from ChIP-seq peaks with DHS regions.

Generating BED-formated files from GEO SRA raw ChIP-seq data

The raw data of the ChIP-seq experiments from GEO were converted to FASTQ-format by means of the SRA Toolkit from NCBI (Sequence Read Archive Submissions Staff). Sequence read alignment was performed with Bowtie with parameters suggested by Arnold et al. (2012) (-v 2 -a -m 100 -S, Langmead et al., 2009). Afterwards peaks were called using MACS on the respective treats and control (Zhang et al., 2008). The default MACS output files comprise information on chromosome, start and stop, summit and intensity of all peaks in BED-format. Basis for alignment and peak calling were mouse genome assembly mm9 and human genome assembly hg19, respectively.

Plotting and further calculations

were performed with the help of R statistical programming language or Microsoft Excel.

Analysis for enrichment of Gene Ontology terms

For the enrichment analysis of Gene Ontology terms the DAVID Bioinformatics Resources 'Functional Annotation' tool was used (Huang da et al., 2009). Entrez IDs of gene sets were uploaded and tested against a background of the union of all target lists from the 15 ChIP-seq experiments, comprising 12,344 REST targets.

Jaccard-index

The Jaccard-index is a measure of similarity between two finite samples. It is the fraction of the size of the intersection and the size of the union of the samples (Jaccard, 1901).

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

Data sources

Table S1: The data for each experiment can be downloaded from the mentioned resources. The Gene Expression Omnibus (GEO) can be accessed on `http://www.ncbi.nlm.nih.gov/geo/` (Edgar et al., 2002) and the UCSC Genome Browser on `http://genome.ucsc.edu/`.

Section and data type	Source	Reference
RefSeq genes hg19	UCSC »Downloads »Human »Feb.2009(hg19) »Annotation database »refGene.txt	(Kent et al., 2002)
RefSeq genes mm9	UCSC »Downloads »Mouse »Jul.2007(mm9) »Annotation database »refGene.txt	(Kent et al., 2002)
2.2.1: ChIP-seq	GEO: GSE27148	(Arnold et al., 2012)
2.2.1: Differential mRNA expression	GEO: GSE27114	(Arnold et al., 2012)
2.2.1: ChIP-seq and direct AR targets	supplementary data of the original publication	(Zhu et al., 2012)
3.2.1: ChIP-seq	Myers - Hudson Alpha Institute for Biotechnology 'wgEncodeHaibTfbs/'*	(The Encode Project Consortium I)
3.2.5: small RNA-seq	Gingeras - Cold Spring Harbor Laboratory 'wgEncodeCshlSortRnaSeq/'*	(The Encode Project Consortium I)
3.2.6: DHS	Crawford - Duke University 'wgEncodeOpenChromDnase/'*	(The Encode Project Consortium I)
3.2.1: miRNA-gene pairs	`http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=vert_61` (Predicted Conserved Targets Info)	(Lewis et al., 2003)

*x in `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/x`

S2 Supplementary Data

Supplementary File 1: Potentially REST-regulated genes according to 15 ChIP-seq datasets: (1) gene is present in the cell type, (0) gene is not present in the cell type.

Supplementary File 2: Bonferroni corrected p-values of enrichment of miRNA targets in REST bound genes.

Supplementary File 3: Over-represented miRNAs and corresponding miRNA precursors with ChIP signal in proximity (10 kb) for 39 REST ChIP-seq samples.

Supplementary File 4: 3 mouse gene lists derived from peaks (i) unfiltered (ranked peak-gene association from ChIP-seq), (ii) filtered by expression (peaks close to genes with significant change in expression level) and (iii) filtered by motif (peaks with RE1 binding motif in the sequence).

Supplementary File 5: Subset of TargetScanHuman predictions filtered for over-represented miRNAs families and REST targets.

Supplementary Folder 1: mBISON implemented in Django, R and Perl.

S3 Supplementary Tables

Table S2: REST-regulated miRNAs from Johnson and Buckley (2009). All miRNAs are referred to as human miRNAs.

microRNA	Publication
miR-9, -124, -132	(Conaco et al., 2006)
miR-1d, miR-330	(Johnson et al., 2008b)
miR-7-2, -7-3, -129-2, -137, -146b, -147, -184, -203, -204, -328, -375, -422a, -602, -637, -940, -1179, -1208, -1224, -1249, -1253, -1255a, -1257, -1267, -1301	(Johnson and Buckley, 2009)
miR-9*	(Packer et al., 2008)
miR-29ab, -95, -135b, -139, -153, -212, -218, -346, -455	(Wu and Xie, 2006)

Table S3: Pairs of miRNA families have seeds that share at least 6 nucleotides resulting in overlapping lists of target genes from TargetScanHuman 6.2.

miRNA family 1	miRNA family 2
let-7/98/4458/4500	miR-202-3p
let-7/98/4458/4500	miR-196abc
miR-208ab/208ab-3p	miR-499-5p
miR-153	miR-448/448-3p
miR-101/101ab	miR-144
miR-17/17-5p/20ab/20b-5p/93/106ab/427/518a-3p/519d	miR-93/93a/105/106a/291a-3p/294/295/302abcde/372/373/428/519a/520be/520acd-3p/1378/1420ac

Table S4: Overview on peak calling algorithms used in ChIP-seq experiments. The summary is based on information from Wilbanks and Facciotti (2010).

Peak caller	reference	properties
MACS	(Zhang et al., 2008)	S, significance calculated in sliding window using a local Poisson model as compared to control or Poisson background model
QuEST	(Valouev et al., 2008)	S, calculates q-value rank using Poisson model, demands control sample
Useq	(Nix et al., 2008)	S, significance calculated in sliding window using a conditional binomial model as compared to Poisson background model or after background subtraction when control is supplied
Minimal ChIPSeq Peak-Finder	(Johnson et al., 2007)	E, peak candidate if more than x reads separated by not more than 100 bps and > 5 overlapping reads, fold enrichment over normalized control, advanced to E-RANGE algorithm
FindPeaks	(Fejes et al., 2008)	E, overlapping XSETs are merged, directionality is used to filter reads that appear to be 'noise', FDR is calculated based on Monte Carlo simulation
PeakSeq	(Rozowsky et al., 2009)	E, enrichment of overlap count/nucleotide compared to a simulated null background model identifies candidate peaks that are compared to by linear regression normalized control using a conditional binomial model to assess significance
Sole-Search	(Blahnik et al., 2010)	E, identifies large deletions or duplications, use background model of sequencable tags to find statistically significant height cutoff for peaks and compare by t-distribution to control
CisGenome	(Ji et al., 2008)	candidates are regions with read counts higher than user-defined cutoff in a sliding window, assesses significance using a negative binomial background model or a conditional binomial model when a control is supplied, use read directionality to filter reads that appear to be 'noise'

S reads are shifted into 3' direction
E reads are extended to the estimated size of DNA fragments

Table S4 – *Continued on next page*

Table S4 – *Continued from previous page*

Peak caller	reference	properties
HPeak	(Qin et al., 2010)	E, genome is segmented into bins, counts XSETs in bins and uses two-state Hidden Markow Model to distinguish peaks from a modeled background, a control can be used to support background modeling
SISSRS	(Jothi et al., 2008)	E, genome is segmented into bins, read count per bin is checked, peaks are where the majority of reads switches between the strands, significance is assessed using a control or Poisson background model
CSDeconv	(Lun et al., 2009)	applies Gaussian kernel density estimator and directionality of reads together with deconvolution, but intense computational demands, suitable for microbial ChIP-seq
PeakRanger	(Feng et al., 2011)	E, broad candidate regions found by thresholding as in PeakSeq, summit-valley-alternator used to detect peak summits
JAMM	(Ibrahim et al., 2015)	splits genome into bins (size determined by heuristic function), enriched bins are joint to broad or narrow peaks, takes covariance of multiple replicates into account
DFilter	(Kumar et al., 2013)	S, genome is segmented into bins and read count is normalized by means of control sample, a training set of positive regions is designed and used to build a linear detection filter as function of mean and covariance, this is used to smoothen the curve, afterwards peaks are detected by thresholding

S reads are shifted into 3' direction
E reads are extended to the estimated size of DNA fragments

Table S5: Comparison of peak-gene association methods in terms of precision and sensitivity in mouse NPs.

cell type	peak-gene association method	nr. of found genes	total nr. of TP genes	nr. of found TPs	precision [%]	sensitivity [%]
NP	binary-10kb	361	379	93	25.8	24.5
NP	binary-1kb	116	379	59	50.9	15.6
NP	binary-20kb	670	379	109	16.3	28.8
NP	binary-2kb	161	379	75	46.6	19.8
NP	binary-50kb	1979	379	140	7.1	36.9
NP	binary-5kb	235	379	84	35.7	22.2
NP	ClostestGene-1	1483	379	133	9.0	35.1
NP	ClostestGene-2	541	379	102	18.9	26.9
NP	ClostestGene-3	182	379	55	30.2	14.5
NP	ClostestGene-4	73	379	26	35.6	6.9
NP	ClostestGene-5	32	379	10	31.2	2.6
NP	ClostestGene-6	21	379	6	28.6	1.6
NP	ClostestGene-7	11	379	1	9.1	0.3
NP	linear-0.5	884	379	111	12.6	29.3
NP	linear-0.6	725	379	111	15.3	29.3
NP	linear-0.7	571	379	105	18.4	27.7
NP	linear-0.8	427	379	97	22.7	25.6
NP	linear-0.9	313	379	93	29.7	24.5
NP	linear-1	85	379	22	25.9	5.8
NP	Ouyang-0.1	340	379	91	26.8	24.0
NP	Ouyang-10	26	379	15	57.7	4.0
NP	Ouyang-10	224	379	82	36.6	21.6
NP	Ouyang-5	111	379	51	45.9	13.5
NP	nonrank	437	379	99	22.7	26.1
NP	rank	405	379	94	23.2	24.8
NP	strict	358	379	91	25.4	24.0
NP	TIP-0.05	32	379	9	28.1	2.4
NP	TIP-0.1	41	379	13	31.7	3.4

TP = true positive

Table S6: The search for enrichment miRNAs was repeated with a TargetScanHuman 6.2 dataset, of which all genes with overlapping seed regions for miR-448 and miR-153 had been removed. The table presents results for miR-448 with FDRs < 0.2.

Cell type	FDR
GM12878	0.008
H1-hESC	0.005
HCT-116	0.191
HeLa-S3	0.008
HL-60	0.199
MCF-7	0.168

Table S7: List of miRNAs with implication as tumor suppressor in glioblastoma and the corresponding source of literature.

miRNA	Reference
miR-107	(Kefas et al., 2009)
miR-124	(Lv and Yang, 2013)
miR-124, miR-137	(Silber et al., 2008)
miR-143	(Zhao et al., 2013b)
miR-145	(Rani et al., 2013)
miR-145, miR-136, miR-129, miR-342, miR-376a	(Haapa-Paananen et al., 2013)
miR-153	(Zhao et al., 2013a)
miR-193-a-3p	(Kwon et al., 2013)
miR-203	(He et al., 2013a)
miR-219-5p	(Rao et al., 2013)
miR-29c	(Wang et al., 2013b)
miR-326	(Wang et al., 2013a)
miR-326, miR-130a	(Qiu et al., 2013)
miR-34a, miR-100, miR-106a, miR-135a, miR-136, miR-181abd, miR-195, miR-205, miR-218, miR-451	(Palumbo et al., 2013)
miR-383	(He et al., 2013b)
miR-491-5p	(Li et al., 2015)
miR-7	(Fang et al., 2012)
miR-708	(Guo et al., 2013)

Table S8: 21 miRNA families with neural specific expression pattern according to TSmiR (Guo et al., 2014). A family was regarded as specifically expressed in neural tissue when one of the family members was listed with this property in TSmiR.

miRNA family
miR-124/124ab/506
miR-125a-3p/1554
miR-125a-5p/125b-5p/351/670/4319
miR-128/128ab
miR-129-5p/129ab-5p
miR-132/212/212-3p
miR-137/137ab
miR-138/138ab
miR-143/1721/4770
miR-149
miR-153
miR-186
miR-199ab-5p
miR-214/761/3619-5p
miR-31
miR-326/330/330-5p
miR-346
miR-7/7ab
miR-93/93a/105/106a/291a-3p/294/295/302abcde/372/373/428/519a/520be/520acd-3p/1378/1420ac
miR-9/9ab
miR-99ab/100

Table S9: Total numbers and percentages of genes with REST peaks contained in a DHS site (DHS-REST) and total number of genes for eight ENCODE cell lines.

Cell line	Genes in DHS-REST region	REST target genes	Fraction REST genes in DHS region [%]
A549	7174	8356	85.9
ECC1	1364	1860	73.3
GM12878	949	1928	49.2
H1-hESC	1859	2919	63.69
HepG2	1943	2639	73.63
K562	3557	4230	84.09
MCF-7	1004	1330	75.49
SK-N-SH	4489	6734	66.66

Table S10: Peak-gene association was conducted by means of DHS regions, which contained a RE1 binding site. The table shows enrichment miRNAs with a FDR > 0.1.

miRNA	ECC1	GM12878	H1-hESC	HepG2	K562	MCF-7
miR-125	-	0.015	0.006	-	-	-
miR-128	-	-	0.011	-	-	-
miR-129	0.015	-	0.004	-	0.064	0.062
miR-138	0.008	0.008	0.015	-	0.005	-
miR-149	-	-	0.055	-	0.068	-
miR-153	-	0.018	0.008	-	0.001	-
miR-214	-	-	0.026	-	-	-
miR-24	-	-	0.052	-	-	-
miR-326	-	0.026	0.004	-	0.004	-
miR-34	-	-	0.012	-	0.015	-
miR-448	-	0.008	0.004	-	0.005	-

S4 Supplementary Figures

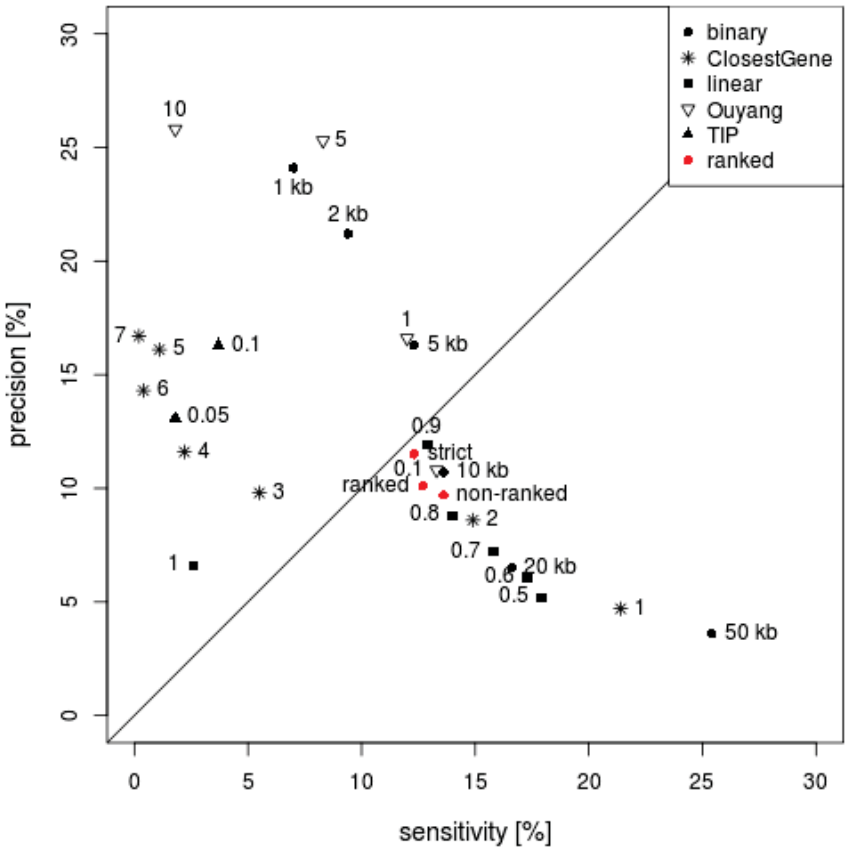

Figure S1: Comparison of peak-gene association methods. Precision is plotted against sensitivity. Genes were assigned to ChIP-seq peaks using the example of REST ESCs. 457 genes were up-regulated after knockout of REST in respect to the wild type (point of reference for sensitivity).
Data point labels: Binary - targets in range of 1 to 50 kb window size. ClosestGene - targets with score higher than 1, 2, 3, 4, 5, 6 and 7. Linear - targets with score higher than 0.5, 0.6, 0.7, 0.8, 0.9 and 1. Ouyang - targets with score higher than 0.1, 0.5, 5 and 10. TIP - targets with p-value smaller than 0.05 and 0.1.

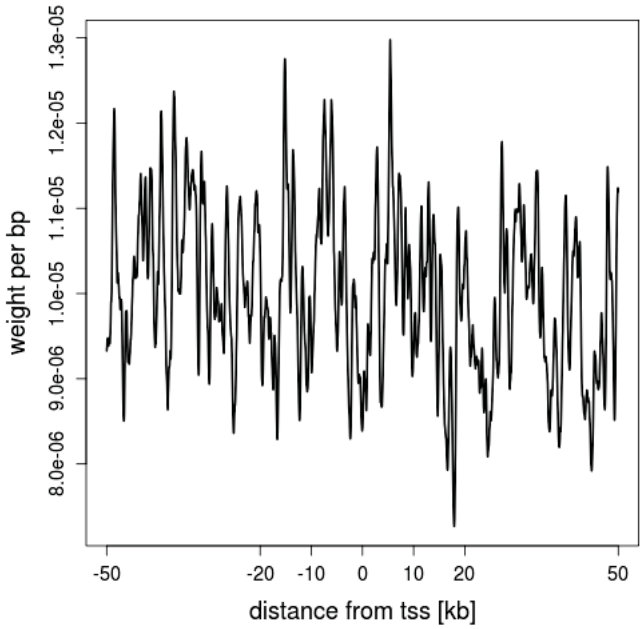

Figure S2: Androgen Receptor binding profile in a prostate cancer model within a range of +/- 50 kb. Weights were generated by the TIP algorithm from ChIP-seq data.

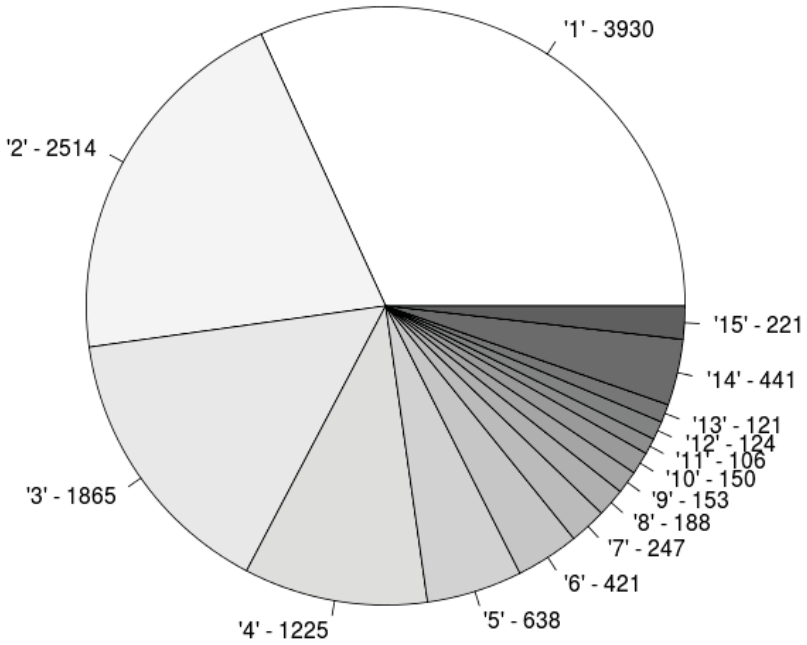

Figure S3: 12,344 REST targets classified according to the number of cell types in which they were detected.

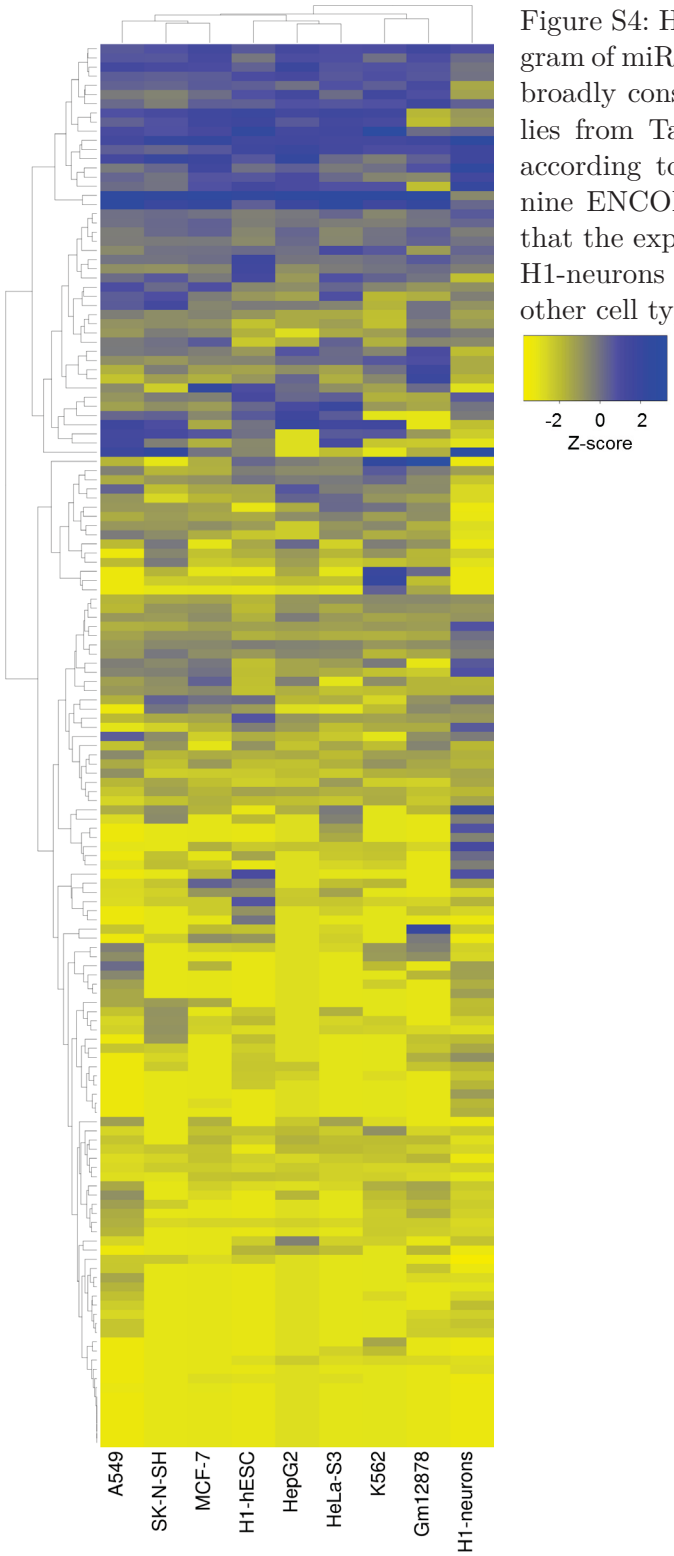

Figure S4: Heatmap and dendrogram of miRNA expression of 153 broadly conserved miRNA families from TargetScanHuman 6.2 according to small RNA-seq on nine ENCODE cell types shows that the expression profile of the H1-neurons differs a lot from the other cell types.

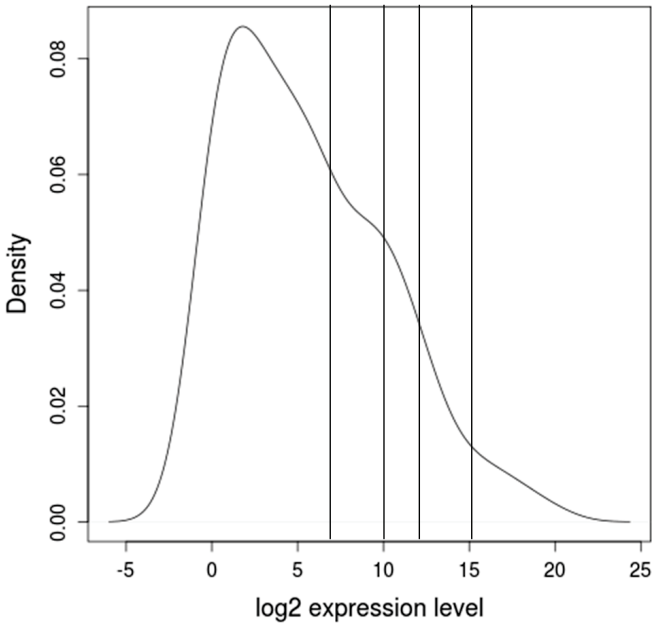

Figure S5: Density distribution of expression levels of 153 miRNAs in H1-neurons. Thresholds used for assessment of over-representation of enrichment miRNAs in H1-neuron expressed miRNAs (7,10,12,15) are marked with vertical lines.

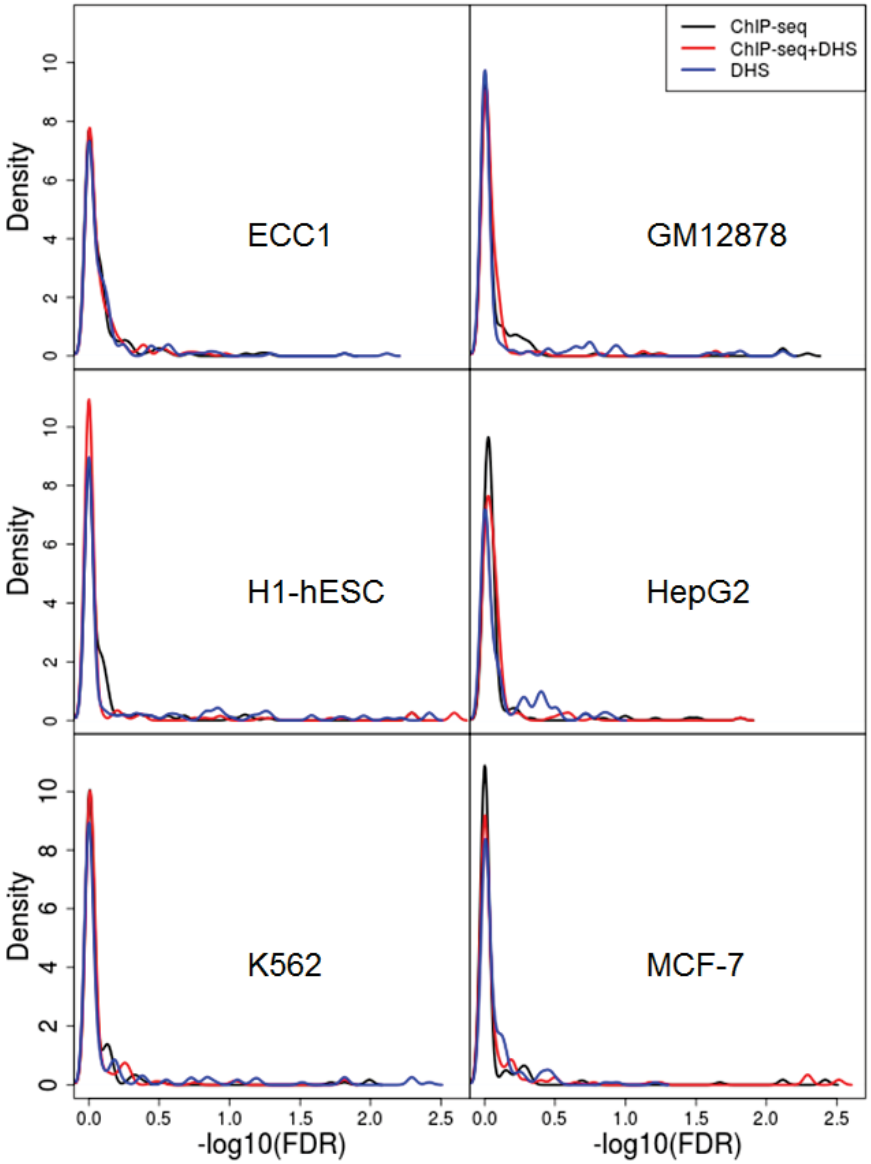

Figure S6: Density distribution of -$\log_{10}$FDRs of 153 miRNAs in six cell types. Basis for the search for enrichment miRNAs were three gene lists:
1. From ChIP-seq data.
2. From regions of overlapping ChIP and DHS signal.
3. From regions of open chromatin with RE1 motif.

Bibliography

Abdi, H. The Bonferroni and Šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 2007.

Alon, U. Network motifs: theory and experimental approaches. *Nat Rev Genet*, 8(6): 450–61, 2007.

Andres, M. E., Burger, C., Peral-Rubio, M. J., Battaglioli, E., Anderson, M. E., Grimes, J., Dallman, J., Ballas, N., and Mandel, G. CoREST: a functional corepressor required for regulation of neural-specific gene expression. *Proc Natl Acad Sci U S A*, 96(17): 9873–8, 1999.

Arnold, P., Scholer, A., Pachkov, M., Balwierz, P. J., Jorgensen, H., Stadler, M. B., van Nimwegen, E., and Schubeler, D. Modeling of epigenome dynamics identifies transcription factors that mediate polycomb targeting. *Genome Res*, 2012.

Bartel, D. P. microRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2): 281–97, 2004.

Bartel, D. P. microRNAs: target recognition and regulatory functions. *Cell*, 136(2): 215–33, 2009.

Beak, D., Villén, J., Shin, C., Camargo, F. D., Gygi, S. P., and Bartel, D. P. The impact of microRNAs on protein output. *Nature*, 455:64–71, 2008.

Belmont, A. S. Large-scale chromatin organization: the good, the surprising, and the still perplexing. *Curr Opin Cell Biol*, 26:69–78, 2014.

Belton, J. M. Hi-c: a comprehensive technique to capture the conformation of genomes. *Methods*, 58(3):268–76, 2012.

Benjamini, Y. and Hochbert, Y. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society*, 57:289–300, 1995.

Blahnik, K. R., Dou, L., O'Geen, H., McPhillips, T., Xu, X., Cao, A. R., Iyengar, S., Nicolet, C. M., Ludascher, B., Korf, I., and Farnham, P. J. Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. *Nucleic Acids Res*, 38(3):e13, 2010.

Boeva, V., Lermine, A., Barette, C., Guillouf, C., and Barillot, E. Nebula-a web-server for advanced ChIP-seq data analysis. *Bioinformatics*, 28(19):2517–9, 2012.

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., Gifford, D. K., Melton, D. A., Jaenisch, R., and Young, R. A. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6):947–56, 2005.

Bruce, A. W., Donaldson, I. J., Wood, I. C., Yerbury, S. A., Sadowski, M. I., Chapman, M., Gottgens, B., and Buckley, N. J. Genome-wide analysis of repressor element 1 silencing transcription factor/neuron-restrictive silencing factor (REST/NRSF) target genes. *Proc Natl Acad Sci U S A*, 101(28):10458–63, 2004.

Bruce, A. W., López C., A., Flicek, P., Down, T., Dhami, P., Dillon, S., Koch, C., Langford, C., Dunham, I., Andrews, R., and Vetrie, D. Functional diversity for REST (NRSF) is defined by *in vivo* binding affinity hierarchies at the DNA sequence level. *Genome research*, 19(6):994–1005, 2009.

Central Brain Tumor Registry of the United States (CBTRUS), . Cbtrus statistical report: Primary brain and central nervous system tumors diagnosed in the united states in 2004-2007. *http://www.cbtrus.org/2011-NPCR-SEER/WEB-0407-Report-3-3-2011.pdf*, 2011.

Chandra, V., Girijadevi, R., Nair, A. S., Pillai, S. S., and Pillai, R. M. MTar: a computational microRNA target prediction architecture for human transcriptome. *BMC Bioinformatics*, 11 Suppl 1:S2, 2010.

Chen, Z.-F., Paquette, A. J., and Anderson, D. J. NRSF/REST is required *in vivo* for repression of multiple neuronal target genes during embryogenesis. *Nature Genetics*, 20:7, 1998.

Cheng, C., Min, R., and Gerstein, M. TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics*, 27(23): 3221–7, 2011.

Chong, J. A., Tapia-Ramirez, J., Kim, S., Toledo-Aral, J. J., Zheng, Y., Boutros, M. C., Altshuller, Y. M., Frohman, M. A., Kraner, S. D., and Mandel, G. REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell*, 80(6):949–57, 1995.

Conaco, C., Otto, S., Han, J. J., and Mandel, G. Reciprocal actions of REST and a microRNA promote neuronal identity. *Proc Natl Acad Sci U S A*, 103(7):2422–7, 2006.

Conti, L., Crisafulli, L., Caldera, V., Tortoreto, M., Brilli, E., Conforti, P., Zunino, F., Magrassi, L., Schiffer, D., and Cattaneo, E. REST controls self-renewal and tumorigenic competence of human glioblastoma cells. *PLoS One*, 7(6):e38486, 2012.

Coulson, J. M., Edgson, J. L., Woll, P. J., and Quinn, J. P. A splice variant of the neuron-restrictive silencer factor repressor is expressed in small cell lung cancer: a potential role in derepression of neuroendocrine genes and a useful clinical marker. *Cancer Res*, 60(7):1840–4, 2000.

Datta, M. and Bhattacharyya, N. P. Regulation of RE1 protein silencing transcription factor (REST) expression by HIP1 protein interactor (HIPPI). *J Biol Chem*, 286(39): 33759–69, 2011.

Dietrich, N., Lerdrup, M., Landt, E., Agrawal-Singh, S., Bak, M., Tommerup, N., Rappsilber, J., Sodersten, E., and Hansen, K. REST-mediated recruitment of polycomb repressor complexes in mammalian cells. *PLoS Genet*, 8(3):e1002494, 2012.

Django Version 1.5, . Retrieved from `https://djangoproject.com`. *Computer Software, Lawrence, Kansas*, 2013.

Doxakis, E. Post-transcriptional regulation of alpha-synuclein expression by mir-7 and mir-153. *J Biol Chem*, 285(17):12726–34, 2010.

Dweep, H., Sticht, C., Pandey, P., and Gretz, N. miRWalk - database: Prediction of possible miRNA binding sites by "walking" the genes of three genomes. *J Biomed Inform*, 2011.

Ebert, M. S. and Sharp, P. A. Roles for microRNAs in conferring robustness to biological processes. *Cell*, 149(3):515–24, 2012.

Edgar, R., Domrachev, M., and Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–10, 2002.

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D. S. microRNA targets in *Drosophila*. *Genome Biol*, 5(1):R1, 2003.

Fang, Q., Strand, A., Law, W., Faca, V. M., Fitzgibbon, M. P., Hamel, N., Houle, B., Liu, X., May, D. H., Poschmann, G., Roy, L., Stuhler, K., Ying, W., Zhang, J., Zheng, Z., Bergeron, J. J., Hanash, S., He, F., Leavitt, B. R., Meyer, H. E., Qian, X., and McIntosh, M. W. Brain-specific proteins decline in the cerebrospinal fluid of humans with Huntington disease. *Mol Cell Proteomics*, 8(3):451–66, 2009.

Fang, Y., Xue, J. L., Shen, Q., Chen, J., and Tian, L. microRNA-7 inhibits tumor growth and metastasis by targeting the phosphoinositide 3-kinase/Akt pathway in hepatocellular carcinoma. *Hepatology*, 55(6):1852–62, 2012.

Fejes, A. P., Robertson, G., Bilenky, M., Varhol, R., Bainbridge, M., and Jones, S. J. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15):1729–30, 2008.

Feng, X., Grossman, R., and Stein, L. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics*, 12:139, 2011.

Friedman, R. C., Farh, K. K., Burge, C. B., and Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19(1):92–105, 2009.

Fuller, G. N., Su, X., Price, R. E., Cohen, Z. R., Lang, F. F., Sawaya, R., and Majumder, S. Many human medulloblastoma tumors overexpress repressor element-1 silencing transcription (REST)/neuron-restrictive silencer factor, which can be functionally countered by REST-VP16. *Mol Cancer Ther*, 4(3):343–9, 2005.

Gao, Z., Ding, P., and Hsieh, J. Profiling of REST-dependent microRNAs reveals dynamic modes of expression. *Front Neurosci*, 6:67, 2012.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–15, 2004.

Gebhardt, M. L., Reuter, S., Mrowka, R., and Andrade-Navarro, M. A. Similarity in targets with REST points to neural and glioblastoma related miRNAs. *Nucleic Acids Res*, 42(9):5436–46, 2014.

Gebhardt, M. L., Mer, A. S., and Andrade-Navarro, M. A. mBISON: Finding miRNA target over-representation in gene lists from ChIP-sequencing data. *BMC Res Notes*, 8:157, 2015.

Gillies, S. G., Haddley, K., Vasiliou, S. A., Jacobson, G. M., von Mentzer, B., Bubb, V. J., and Quinn, J. P. Distinct gene expression profiles directed by the isoforms of the transcription factor neuron-restrictive silencer factor in human SK-N-AS neuroblastoma cells. *J Mol Neurosci*, 44(2):77–90, 2011.

Grimson, A., Farh, K. K., Johnston, W. K., Garrett-Engele, P., Lim, L. P., and Bartel, D. P. microRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, 27(1):91–105, 2007.

Grosswendt, S., Filipchyk, A., Manzano, M., Klironomos, F., Schilling, M., Herzog, M., Gottwein, E., and Rajewsky, N. Unambiguous identification of miRNA:target site interactions by different types of ligation reactions. *Mol Cell*, 54(6):1042–54, 2014.

Guo, P., Lan, J., Ge, J., Nie, Q., Mao, Q., and Qiu, Y. miR-708 acts as a tumor suppressor in human glioblastoma cells. *Oncol Rep*, 30(2):870–6, 2013.

Guo, Z., Maki, M., Ding, R., Yang, Y., Zhang, B., and Xiong, L. Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues. *Sci Rep*, 4:5150, 2014.

Gupta, R., Bhattacharyya, A., Agosto-Perez, F. J., Wickramasinghe, P., and Davuluri, R. V. MPromDb update 2010: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-seq experimental data. *Nucleic Acids Res*, 39(Database issue):D92–7, 2011.

Ha, M. and Kim, V. N. Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol*, 15 (8):509–24, 2014.

Haapa-Paananen, S., Chen, P., Hellstrom, K., Kohonen, P., Hautaniemi, S., Kallioniemi, O., and Perala, M. Functional profiling of precursor microRNAs identifies microRNAs essential for glioma proliferation. *PLoS One*, 8(4):e60930, 2013.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, J. M., Jungkamp, A. C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–41, 2010.

Hampf, M. and Gossen, M. A protocol for combined photinus and renilla luciferase quantification compatible with protein assays. *Anal Biochem*, 356(1):94–9, 2006.

He, J., Deng, Y., Yang, G., and Xie, W. microRNA-203 down-regulation is associated with unfavorable prognosis in human glioma. *J Surg Oncol*, 108(2):121–5, 2013a.

He, X., Chen, C. C., Hong, F., Fang, F., Sinha, S., Ng, H. H., and Zhong, S. A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PLoS One*, 4(12):e8155, 2009.

He, Z., Cen, D., Luo, X., Li, D., Li, P., Liang, L., and Meng, Z. Downregulation of miR-383 promotes glioma cell invasion by targeting insulin-like growth factor 1 receptor. *Med Oncol*, 30(2):557, 2013b.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell*, 38(4):576–89, 2010.

Hsu, J. B., Chiu, C. M., Hsu, S. D., Huang, W. Y., Chien, C. H., Lee, T. Y., and Huang, H. D. miRTar: an integrated system for identifying miRNA-target interactions in human. *BMC Bioinformatics*, 12(1):300, 2011.

Huang, Y., Myers, S. J., and Dingledine, R. Transcriptional repression by REST: recruitment of Sin3A and histone deacetylase to neuronal genes. *Nat Neurosci*, 2(10): 867–72, 1999.

Huang da, W., Sherman, B. T., and Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4(1):44–57, 2009.

Huntzinger, E. and Izaurralde, E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet*, 12(2):99–110, 2011.

Hwang, J. Y., Kaneko, N., Noh, K. M., Pontarelli, F., and Zukin, R. S. The gene silencing transcription factor REST represses miR-132 expression in hippocampal neurons destined to die. *J Mol Biol*, 426(20):3454–66, 2014.

Ibrahim, M. M., Lacadie, S. A., and Ohler, U. JAMM: a peak finder for joint analysis of NGS replicates. *Bioinformatics*, 31(1):48–55, 2015.

Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

Ji, H., Jiang, H., Ma, W., Johnson, D. S., Myers, R. M., and Wong, W. H. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol*, 26(11): 1293–300, 2008.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, 316(5830):1497–502, 2007.

Johnson, R. and Buckley, N. J. Gene dysregulation in Huntington's disease: REST, microRNAs and beyond. *Neuromolecular Med*, 11(3):183–99, 2009.

Johnson, R., Gamblin, R. J., Ooi, L., Bruce, A. W., Donaldson, I. J., Westhead, D. R., Wood, I. C., Jackson, R. M., and Buckley, N. J. Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. *Nucleic Acids Res*, 34(14):3862–77, 2006.

Johnson, R., Teh, C. H., Kunarso, G., Wong, K. Y., Srinivasan, G., Cooper, M. L., Volta, M., Chan, S. S., Lipovich, L., Pollard, S. M., Karuturi, R. K., Wei, C. L., Buckley, N. J., and Stanton, L. W. REST regulates distinct transcriptional networks in embryonic and neural stem cells. *PLoS Biol*, 6(10):e256, 2008a.

Johnson, R., Zuccato, C., Belyaev, N. D., Guest, D. J., Cattaneo, E., and Buckley, N. J. A microRNA-based gene dysregulation pathway in Huntington's disease. *Neurobiol Dis*, 29(3):438–45, 2008b.

Johnson, R., Teh, C. H., Jia, H., Vanisri, R. R., Pandey, T., Lu, Z. H., Buckley, N. J., Stanton, L. W., and Lipovich, L. Regulation of neural macroRNAs by the transcriptional repressor REST. *RNA*, 15(1):85–96, 2009.

Jothi, R., Cuddapah, S., Barski, A., Cui, K., and Zhao, K. Genome-wide identification of *in vivo* protein-DNA binding sites from ChIP-seq data. *Nucleic Acids Res*, 36(16): 5221–31, 2008.

Kamal, M. M., Sathyan, P., Singh, S. K., Zinn, P. O., Marisetty, A. L., Liang, S., Gumin, J., El-Mesallamy, H. O., Suki, D., Colman, H., Fuller, G. N., Lang, F. F., and Majumder, S. REST regulates oncogenic properties of glioblastoma stem cells. *Stem Cells*, 30(3):405–14, 2012.

Karolchik, D., Kuhn, R. M., Baertsch, R., Barber, G. P., Clawson, H., Diekhans, M., Giardine, B., Harte, R. A., Hinrichs, A. S., Hsu, F., Kober, K. M., Miller, W., Pedersen, J. S., Pohl, A., Raney, B. J., Rhead, B., Rosenbloom, K. R., Smith, K. E., Stanke, M., Thakkapallayil, A., Trumbower, H., Wang, T., Zweig, A. S., Haussler, D., and Kent, W. J. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res*, 36(Database issue):D773–9, 2008.

Kefas, B., Comeau, L., Floyd, D. H., Seleverstov, O., Godlewski, J., Schmittgen, T., Jiang, J., diPierro, C. G., Li, Y., Chiocca, E. A., Lee, J., Fine, H., Abounader, R., Lawler, S., and Purow, B. The neuronal microRNA miR-326 acts in a feedback loop with Notch and has therapeutic potential against brain tumors. *J Neurosci*, 29(48): 15161–8, 2009.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. The human genome browser at UCSC. *Genome Res*, 12(6):996–1006, 2002.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. The role of site accessibility in microRNA target recognition. *Nat Genet*, 39(10):1278–84, 2007.

Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol*, 26(12):1351–9, 2008.

Kiriakidou, M., Nelson, P. T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z., and Hatzigeorgiou, A. A combined computational-experimental approach predicts human microRNA targets. *Genes Dev*, 18(10):1165–78, 2004.

Kozomara, A. and Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*, 39(Database issue):D152–7, 2011.

Kraner, S. D., Chong, J. A., Tsay, H. J., and Mandel, G. Silencing the type II sodium channel gene: a model for neural-specific gene regulation. *Neuron*, 9(1):37–44, 1992.

Krek, A., Grun, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M., and Rajewsky, N. Combinatorial microRNA target predictions. *Nat Genet*, 37(5):495–500, 2005.

Kumar, V., Muratani, M., Rayan, N. A., Kraus, P., Lufkin, T., Ng, H. H., and Prabhakar, S. Uniform, optimal signal processing of mapped deep-sequencing data. *Nat Biotechnol*, 31(7):615–22, 2013.

Kuwahara, K. Role of NRSF/REST in the regulation of cardiac gene expression and function. *Circ J*, 77(11):2682–6, 2013.

Kwon, J. E., Kim, B. Y., Kwak, S. Y., Bae, I. H., and Han, Y. H. Ionizing radiation-inducible microRNA miR-193a-3p induces apoptosis by directly targeting Mcl-1. *Apoptosis*, 18(7):896–909, 2013.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A. O., Landthaler, M., Lin, C., Socci, N. D., Hermida, L., Fulci, V., Chiaretti, S., Foa, R., Schliwka, J., Fuchs, U., Novosel, A., Muller, R. U., Schermer, B., Bissels, U., Inman, J., Phan, Q., Chien, M., Weir, D. B., Choksi, R., De Vita, G., Frezzetti, D., Trompeter, H. I., Hornung, V., Teng, G., Hartmann, G., Palkovits, M., Di Lauro, R., Wernet, P., Macino, G., Rogler, C. E., Nagle, J. W., Ju, J., Papavasiliou, F. N., Benzing, T., Lichter, P., Tam, W., Brownstein, M. J., Bosio, A., Borkhardt,

A., Russo, J. J., Sander, C., Zavolan, M., and Tuschl, T. A mammalian microRNA Expression Atlas based on small RNA library sequencing. *Cell*, 129(7):1401–14, 2007.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. Software for computing and annotating genomic ranges. *PLoS Comput Biol*, 9(8):e1003118, 2013.

Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., and de Graaff, E. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet*, 12(14):1725–35, 2003.

Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–98, 2003.

Lewis, B. P., Burge, C. B., and Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120 (1):15–20, 2005.

Li, H., Ruan, J., and Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–8, 2008.

Li, X., Liu, Y., Granberg, K. J., Wang, Q., Moore, L. M., Ji, P., Gumin, J., Sulman, E. P., Calin, G. A., Haapasalo, H., Nykter, M., Shmulevich, I., Fuller, G. N., Lang, F. F., and Zhang, W. Two mature products of miR-491 coordinate to suppress key cancer hallmarks in glioblastoma. *Oncogene*, 34(13):1619–28, 2015.

Liang, C., Zhu, H., Xu, Y., Huang, L., Ma, C., Deng, W., Liu, Y., and Qin, C. microRNA-153 negatively regulates the expression of amyloid precursor protein and amyloid precursor-like protein 2. *Brain Res*, 1455:103–13, 2012.

Liang, J., Tong, P., Zhao, W., Li, Y., Zhang, L., Xia, Y., and Yu, Y. The REST gene signature predicts drug sensitivity in neuroblastoma cell lines and is significantly associated with neuroblastoma tumor stage. *Int J Mol Sci*, 15(7):11220–33, 2014.

Liang, Y., Ridzon, D., Wong, L., and Chen, C. Characterization of microRNA expression profiles in normal human tissues. *BMC Genomics*, 8:166, 2007.

Liu, B., Cheng, S., Xing, W., Pourteymoor, S., and Mohan, S. RE1-silencing transcription factor (Rest) is a novel regulator of osteoblast differentiation. *J Cell Biochem*, 2015.

Liu, C., Rennie, W. A., Mallick, B., Kanoria, S., Long, D., Wolenc, A., Carmack, C. S., and Ding, Y. microRNA binding sites in *C. elegans* 3' UTRs. *RNA Biol*, 11(6): 693–701, 2014.

Liu, C. M., Wang, R. Y., Saijilafu, Jiao, Z. X., Zhang, B. Y., and Zhou, F. Q. microRNA-138 and SIRT1 form a mutual negative feedback loop to regulate mammalian axon regeneration. *Genes Dev*, 27(13):1473–83, 2013.

Lu, T., Aron, L., Zullo, J., Pan, Y., Kim, H., Chen, Y., Yang, T. H., Kim, H. M., Drake, D., Liu, X. S., Bennett, D. A., Colaiacovo, M. P., and Yankner, B. A. REST and stress resistance in ageing and Alzheimer's disease. *Nature*, 507(7493):448–54, 2014.

Lun, D. S., Sherrid, A., Weiner, B., Sherman, D. R., and Galagan, J. E. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. *Genome Biol*, 10(12):R142, 2009.

Lv, Z. and Yang, L. miR-124 inhibits the growth of glioblastoma through the downregulation of SOS1. *Mol Med Rep*, 8(2):345–9, 2013.

Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C. Y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., and Wasserman, W. W. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res*, 42(Database issue):D142–7, 2014.

Megraw, M., Mukherjee, S., and Ohler, U. Sustained-input switches for transcription factors and microRNAs are central building blocks of eukaryotic gene circuits. *Genome Biol*, 14(8):R85, 2013.

Mikulak, J., Negrini, S., Klajn, A., D'Alessandro, R., Mavilio, D., and Meldolesi, J. Dual REST-dependence of L1CAM: from gene expression to alternative splicing governed by Nova2 in neural cells. *J Neurochem*, 120(5):699–709, 2012.

Miska, E. A., Alvarez-Saavedra, E., Abbott, A. L., Lau, N. C., Hellman, A. B., McGonagle, S. M., Bartel, D. P., Ambros, V. R., and Horvitz, H. R. Most *Caenorhabditis elegans* microRNAs are individually not essential for development or viability. *PLoS Genet*, 3(12):e215, 2007.

Mori, N., Schoenherr, C., Vandenbergh, D. J., and Anderson, D. J. A common silencer element in the SCG10 and type II Na+ channel genes binds a factor present in nonneuronal cells but not in neuronal cells. *Neuron*, 9(1):45–54, 1992.

Muinos-Gimeno, M., Espinosa-Parrilla, Y., Guidi, M., Kagerbauer, B., Sipila, T., Maron, E., Pettai, K., Kananen, L., Navines, R., Martin-Santos, R., Gratacos, M., Metspalu, A., Hovatta, I., and Estivill, X. Human microRNAs miR-22, miR-138-2, miR-148a, and miR-488 are associated with panic disorder and regulate several anxiety candidate genes and related pathways. *Biol Psychiatry*, 69(6):526–33, 2011.

Naifang, S., Minping, Q., and Minghua, D. Integrative approaches for microRNA target prediction: Combining sequence information and the paired mRNA and miRNA expression profiles. *Curr Bioinform*, 8(1):37–45, 2013.

Nix, D. A., Courdy, S. J., and Boucher, K. M. Empirical methods for controlling false positives and estimating confidence in ChIP-seq peaks. *BMC Bioinformatics*, 9:523, 2008.

Ooi, L. and Wood, I. C. Chromatin crosstalk in development and disease: lessons from REST. *Nat Rev Genet*, 8(7):544–54, 2007.

Otto, S. J., McCorkle, S. R., Hover, J., Conaco, C., Han, J. J., Impey, S., Yochum, G. S., Dunn, J. J., Goodman, R. H., and Mandel, G. A new binding motif for the transcriptional repressor REST uncovers large gene networks devoted to neuronal functions. *J Neurosci*, 27(25):6729–39, 2007.

Oulas, A., Karathanasis, N., Louloupi, A., Iliopoulos, I., Kalantidis, K., and Poirazi, P. A new microRNA target prediction tool identifies a novel interaction of a putative miRNA with CCND2. *RNA Biol*, 9(9), 2012.

Ouyang, Z., Zhou, Q., and Wong, W. H. ChIP-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci U S A*, 106(51):21521–6, 2009.

Packer, A. N., Xing, Y., Harper, S. Q., Jones, L., and Davidson, B. L. The bifunctional microRNA miR-9/miR-9* regulates REST and CoREST and is downregulated in Huntington's disease. *J Neurosci*, 28(53):14341–6, 2008.

Palm, K., Belluardo, N., Metsis, M., and Timmusk, T. Neuronal expression of zinc finger transcription factor REST/NRSF/XBR gene. *J Neurosci*, 18(4):1280–96, 1998.

Palm, K., Metsis, M., and Timmusk, T. Neuron-specific splicing of zinc finger transcription factor REST/NRSF/XBR is frequent in neuroblastomas and conserved in human, mouse and rat. *Brain Res Mol Brain Res*, 72(1):30–9, 1999.

Palumbo, S., Miracco, C., Pirtoli, L., and Comincini, S. Emerging roles of microRNA in modulating cell-death processes in malignant glioma. *J Cell Physiol*, 2013.

Pan, C., Kumar, C., Bohl, S., Klingmueller, U., and Mann, M. Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type-specific functions. *Mol Cell Proteomics*, 8(3):443–50, 2009.

Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10):669–80, 2009.

Park, S. Y., Kim, J. B., and Han, Y. M. REST is a key regulator in brain-specific homeobox gene expression during neuronal differentiation. *J Neurochem*, 103(6):2565–74, 2007.

Qin, Z. S., Yu, J., Shen, J., Maher, C. A., Hu, M., Kalyana-Sundaram, S., Yu, J., and Chinnaiyan, A. M. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-seq data. *BMC Bioinformatics*, 11:369, 2010.

Qiu, S., Lin, S., Hu, D., Feng, Y., Tan, Y., and Peng, Y. Interactions of miR-323/miR-326/miR-329 and miR-130a/miR-155/miR-210 as prognostic indicators for clinical outcome of glioblastoma patients. *J Transl Med*, 11:10, 2013.

Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H., and Turner, D. J. A large genome center's improvements to the Illumina sequencing system. *Nat Methods*, 5(12):1005–10, 2008.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., and Gu, Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13:341, 2012.

R Development Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria, 2008.

Raj, B., O'Hanlon, D., Vessey, J. P., Pan, Q., Ray, D., Buckley, N. J., Miller, F. D., and Blencowe, B. J. Cross-regulation between an alternative splicing activator and a transcription repressor controls neurogenesis. *Mol Cell*, 43(5):843–50, 2011.

Rani, S. B., Rathod, S. S., Karthik, S., Kaur, N., Muzumdar, D., and Shiras, A. S. miR-145 functions as a tumor-suppressive RNA by targeting Sox9 and adducin 3 in human glioma cells. *Neuro Oncol*, 15(10):1302–16, 2013.

Rao, S. A., Arimappamagan, A., Pandey, P., Santosh, V., Hegde, A. S., Chandramouli, B. A., and Somasundaram, K. miR-219-5p inhibits receptor tyrosine kinase pathway by targeting EGFR in glioblastoma. *PLoS One*, 8(5):e63164, 2013.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., and Young, R. A. Genome-wide location and function of DNA binding proteins. *Science*, 290:2306–09, 2000.

Rennie, W., Liu, C., Carmack, C. S., Wolenc, A., Kanoria, S., Lu, J., Long, D., and Ding, Y. STarMir: a web server for prediction of microRNA binding sites. *Nucleic Acids Res*, 42(Web Server issue):W114–8, 2014.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, 43(7):e47, 2015.

Rodelsperger, C., Guo, G., Kolanczyk, M., Pletschacher, A., Kohler, S., Bauer, S., Schulz, M. H., and Robinson, P. N. Integrative analysis of genomic, functional and protein interaction data predicts long-range enhancer-target gene interactions. *Nucleic Acids Res*, 39(7):2492–502, 2011.

Ronen, M., Rosenberg, R., Shraiman, B. I., and Alon, U. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc Natl Acad Sci U S A*, 99(16):10555–60, 2002.

Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M. B. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol*, 27(1):66–75, 2009.

Sandmann, T., Jensen, L. J., Jakobsen, J. S., Karzynski, M. M., Eichenlaub, M. P., Bork, P., and Furlong, E. E. A temporal map of transcription factor activity: Mef2 directly regulates target genes at all stages of muscle development. *Dev Cell*, 10(6): 797–807, 2006.

Schoenherr, C. J. and Anderson, D. J. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science*, 267(5202):1360–3, 1995.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34(2):166–76, 2003.

Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209): 58–63, 2008.

Sempere, L. F., Freemantle, S., Pitha-Rowe, I., Moss, E., Dmitrovsky, E., and Ambros, V. Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. *Genome Biol*, 5(3):R13, 2004.

Sequence Read Archive Submissions Staff. Using the SRA toolkit to convert .sra files into other formats. *National Center for Biotechnology Information (US)*, 20.05.2015, 2011.

Shalgi, R., Lieber, D., Oren, M., and Pilpel, Y. Global and local architecture of the mammalian microRNA-transcription factor regulatory network. *PLoS Comput Biol*, 3(7):e131, 2007.

Shimojo, M., Lee, J. H., and Hersh, L. B. Role of zinc finger domains of the transcription factor neuron-restrictive silencer factor/repressor element-1 silencing transcription factor in DNA binding and nuclear localization. *J Biol Chem*, 276(16):13121–6, 2001.

Shin, H., Liu, T., Manrai, A. K., and Liu, X. S. CEAS: *cis*-regulatory element annotation system. *Bioinformatics*, 25(19):2605–6, 2009.

Sikora-Wohlfeld, W., Ackermann, M., Christodoulou, E. G., Singaravelu, K., and Beyer, A. Assessing computational methods for transcription factor target gene identification based on ChIP-seq data. *PLoS Comput Biol*, 9(11):e1003342, 2013.

Silber, J., Lim, D. A., Petritsch, C., Persson, A. I., Maunakea, A. K., Yu, M., Vandenberg, S. R., Ginzinger, D. G., James, C. D., Costello, J. F., Bergers, G., Weiss, W. A., Alvarez-Buylla, A., and Hodgson, J. G. miR-124 and miR-137 inhibit proliferation

of glioblastoma multiforme cells and induce differentiation of brain tumor stem cells. *BMC Med*, 6:14, 2008.

Singh, S. K., Kagalwala, M. N., Parker-Thornburg, J., Adams, H., and Majumder, S. REST maintains self-renewal and pluripotency of embryonic stem cells. *Nature*, 453 (7192):223–7, 2008.

Soler, E., Andrieu-Soler, C., de Boer, E., Bryne, J. C., Thongjuea, S., Stadhouders, R., Palstra, R. J., Stevens, M., Kockx, C., van Ijcken, W., Hou, J., Steinhoff, C., Rijkers, E., Lenhard, B., and Grosveld, F. The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes Dev*, 24(3):277–89, 2010.

Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B. K., Sheffield, N. C., Graf, S., Huss, M., Keefe, D., Liu, Z., London, D., McDaniell, R. M., Shibata, Y., Showers, K. A., Simon, J. M., Vales, T., Wang, T., Winter, D., Zhang, Z., Clarke, N. D., Birney, E., Iyer, V. R., Crawford, G. E., Lieb, J. D., and Furey, T. S. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res*, 21(10):1757–67, 2011.

Stark, A., Brennecke, J., Bushati, N., Russell, R. B., and Cohen, S. M. Animal microRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, 123(6):1133–46, 2005.

Steinfeld, I., Navon, R., Ach, R., and Yakhini, Z. miRNA target enrichment analysis reveals directly active miRNAs in health and disease. *Nucleic Acids Res*, 41(3):e45, 2013.

Stempor, P. A., Cauchi, M., and Wilson, P. MMpred: functional miRNA - mRNA interaction analyses by miRNA expression prediction. *BMC Genomics*, 13(1):620, 2012.

Stojnic, R. and Diez, D. PWMEnrich: PWM enrichment analysis. *R package version 4.4.0*, 2014.

Sturm, M., Hackenberg, M., Langenberger, D., and Frishman, D. TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC Bioinformatics*, 11:292, 2010.

Sudarsanam, P., Pilpel, Y., and Church, G. M. Genome-wide co-occurrence of promoter elements reveals a *cis*-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*. *Genome Res*, 12(11):1723–31, 2002.

Taher, L. and Ovcharenko, I. Variable locus length in the human genome leads to ascertainment bias in functional inference for non-coding elements. *Bioinformatics*, 25(5):578–84, 2009.

The Encode Project Consortium I. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol*, 9(4):e1001046, 2011.

The Encode Project Consortium II. ENCODE experiment matrix, `http://genome.ucsc.edu/ENCODE/dataMatrix/encodeDataMatrixHuman.html`, June 11th 2015.

Thiel, G., Lietz, M., and Cramer, M. Biological activity and modular structure of RE-1-silencing transcription factor (REST), a repressor of neuronal genes. *J Biol Chem*, 273(41):26891–9, 1998.

Thomson, D. W., Bracken, C. P., and Goodall, G. J. Experimental strategies for microRNA target identification. *Nucleic Acids Res*, 2011.

Tsang, J., Zhu, J., and van Oudenaarden, A. microRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Mol Cell*, 26(5):753–67, 2007.

Tsang, J. S., Ebert, M. S., and van Oudenaarden, A. Genome-wide dissection of microRNA functions and cotargeting networks using gene set signatures. *Mol Cell*, 38 (1):140–53, 2010.

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M., and Sidow, A. Genome-wide analysis of transcription factor binding sites based on ChIP-seq data. *Nat Methods*, 5(9):829–34, 2008.

Vasudevan, S. Posttranscriptional upregulation by microRNAs. *Wiley Interdiscip Rev RNA*, 3(3):311–30, 2012.

Ventura, A., Young, A. G., Winslow, M. M., Lintault, L., Meissner, A., Erkeland, S. J., Newman, J., Bronson, R. T., Crowley, D., Stone, J. R., Jaenisch, R., Sharp, P. A., and Jacks, T. Targeted deletion reveals essential and overlapping functions of the miR-17 through 92 family of miRNA clusters. *Cell*, 132(5):875–86, 2008.

Vergoulis, T., Vlachos, I. S., Alexiou, P., Georgakilas, G., Maragkakis, M., Reczko, M., Gerangelos, S., Koziris, N., Dalamagas, T., and Hatzigeorgiou, A. G. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res*, 2012.

Vidigal, J. A. and Ventura, A. The biological functions of miRNAs: lessons from *in vivo* studies. *Trends Cell Biol*, 25(3):137–47, 2015.

Visani, M., de Biase, D., Marucci, G., Taccioli, C., Baruzzi, A., Pession, A., and Group, P. S. Definition of miRNAs expression profile in glioblastoma samples: the relevance of non-neoplastic brain reference. *PLoS One*, 8(1):e55314, 2013.

Visvanathan, J., Lee, S., Lee, B., Lee, J. W., and Lee, S. K. The microRNA miR-124 antagonizes the anti-neural REST/SCP1 pathway during embryonic CNS development. *Genes Dev*, 21(7):744–9, 2007.

Wagoner, M. P., Gunsalus, K. T., Schoenike, B., Richardson, A. L., Friedl, A., and Roopra, A. The transcription factor REST is lost in aggressive breast cancer. *PLoS Genet*, 6(6):e1000979, 2010.

Wang, S., Lu, S., Geng, S., Ma, S., Liang, Z., and Jiao, B. Expression and clinical significance of microRNA-326 in human glioma miR-326 expression in glioma. *Med Oncol*, 30(1):373, 2013a.

Wang, Y., Li, Y., Sun, J., Wang, Q., Sun, C., Yan, Y., Yu, L., Cheng, D., An, T., Shi, C., Xu, J., Wei, C., Liu, J., Wen, Y., Zhao, S., Li, H., Zhang, H., Xu, H., and Yu, S. Tumor-suppressive effects of miR-29c on gliomas. *Neuroreport*, 24(12):637–45, 2013b.

Weber, G. L., Parat, M. O., Binder, Z. A., Gallia, G. L., and Riggins, G. J. Abrogation of PIK3CA or PIK3R1 reduces proliferation, migration, and invasion in glioblastoma multiforme cells. *Oncotarget*, 2(11):833–49, 2011.

Wilbanks, E. G. and Facciotti, M. T. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One*, 5(7):e11471, 2010.

Wu, C. and Gilbert, W. Tissue-specific exposure of chromatin structure at the 5' terminus of the rat preproinsulin II gene. *Proc Natl Acad Sci U S A*, 78(3):1577–80, 1981.

Wu, G. and Ji, H. ChIPXpress: using publicly available gene expression data to improve ChIP-seq and ChIP-chip target gene ranking. *BMC Bioinformatics*, 14:188, 2013.

Wu, J. and Xie, X. Comparative sequence analysis reveals an intricate network among REST, CREB and miRNA in mediating neuronal gene expression. *Genome Biol*, 7 (9):R85, 2006.

Yeamans, C., Wang, D., Paz-Priel, I., Torbett, B. E., Tenen, D. G., and Friedman, A. D. C/EBPalpha binds and activates the PU.1 distal enhancer to induce monocyte lineage commitment. *Blood*, 110(9):3136–42, 2007.

Zambelli, F., Pesole, G., and Pavesi, G. PscanChIP: Finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-seq experiments. *Nucleic Acids Res*, 41(Web Server issue):W535–43, 2013.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. Model-based analysis of ChIP-seq (MACS). *Genome Biol*, 9(9):R137, 2008.

Zhao, S., Deng, Y., Liu, Y., Chen, X., Yang, G., Mu, Y., Zhang, D., Kang, J., and Wu, Z. microRNA-153 is tumor suppressive in glioblastoma stem cells. *Mol Biol Rep*, 40 (4):2789–98, 2013a.

Zhao, S., Liu, H., Liu, Y., Wu, J., Wang, C., Hou, X., Chen, X., Yang, G., Zhao, L., Che, H., Bi, Y., Wang, H., Peng, F., and Ai, J. miR-143 inhibits glycolysis and depletes stemness of glioblastoma stem-like cells. *Cancer Lett*, 333(2):253–60, 2013b.

Zhu, L. J., Gazin, C., Lawson, N. D., Pages, H., Lin, S. M., Lapointe, D. S., and Green, M. R. ChIPpeakAnno: a bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*, 11:237, 2010.

Zhu, Z., Shi, M., Hu, W., Estrella, H., Engebretsen, J., Nichols, T., Briere, D., Hosea, N., Los, G., Rejto, P. A., and Fanjul, A. Dose-dependent effects of small-molecule antagonists on the genomic landscape of androgen receptor binding. *BMC Genomics*, 13:355, 2012.

Zuccato, C., Tartari, M., Crotti, A., Goffredo, D., Valenza, M., Conti, L., Cataudella, T., Leavitt, B. R., Hayden, M. R., Timmusk, T., Rigamonti, D., and Cattaneo, E. Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes. *Nat Genet*, 35(1):76–83, 2003.

List of Figures

List of Tables

List of Equations

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Berlin, September 2015 Marie Luise Gebhardt, geb. Sauer